Rajendra Prasath
T. Kathirvalavakumar (Eds.)

Mining Intelligence and Knowledge Exploration

LNAI 8284

First International Conference, MIKE 2013
Tamil Nadu, India, December 2013
Proceedings

Springer

Lecture Notes in Artificial Intelligence 8284

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel
University of Alberta, Edmonton, Canada
Yuzuru Tanaka
Hokkaido University, Sapporo, Japan
Wolfgang Wahlster
DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann
DFKI and Saarland University, Saarbrücken, Germany

Rajendra Prasath T. Kathirvalavakumar (Eds.)

Mining Intelligence and Knowledge Exploration

First International Conference, MIKE 2013
Tamil Nadu, India, December 18-20, 2013
Proceedings

Springer

Volume Editors

Rajendra Prasath
National University of Ireland
University College Cork
Department of Business Information Systems, FSIC
Cork, Ireland
E-mail: drrprasath@gmail.com

T. Kathirvalavakumar
V.H.N.Senthikumara Nadar College (Autonomous)
Research Centre in Computer Science
Virudhunagar - 626 001
Tamil Nadu, India
E-mail: kathirvalavakumar@yahoo.com

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-03843-8 e-ISBN 978-3-319-03844-5
DOI 10.1007/978-3-319-03844-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013954548

CR Subject Classification (1998): I.2, H.3, H.2, I.5, I.4, C.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

© Springer International Publishing Switzerland 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at MIKE 2013: The First International Conference on Mining Intelligence and Knowledge Exploration held during December 18–20, 2013, at the Research Center in Computer Science, Virudhunagar Hindu Nadars' Senthikumara Nadar(VHNSN) College (Autonomous), Tamil Nadu, India (http://www.mike2013.org/). There were 334 submissions from 19 countries. After an initial screening, 289 qualified submissions were selected for review. Each paper was reviewed by at least two Program Committee members using the criteria of relevance, originality, technical quality, and presentation. The committee accepted 82 papers for oral presentation (accpetance rate: 28.3%) and 52 papers for short/poster presentation at the conference.

The International Conference on Mining Intelligence and Knowledge Exploration (MIKE) is an initiative focusing on research and applications on various topics of human intelligence mining and knowledge discovery. The primary goal is to present state-of-art scientific results, to disseminate modern technologies, and to promote collaborative research in mining intelligence and knowledge exploration. Human intelligence has evolved steadily over several generations, and today human expertise is excelling in individual domains and in knowledge-acquiring tasks. Thus, mining human intelligence becomes an essential and exciting part of human expertise/knowledge exploration tasks. The primary objective is to focus on the frontiers of human intelligence mining as a unified research field.

The accepted papers were chosen on the basis of their contribution which, in certain ways, provides a starting point to young researchers involved in investigating the learning algorithms and knowledge-discovery tasks. We have grouped the accepted papers into various subtopics such as feature selection, classification, clustering, image procesing, network security, speech processing, machine learning, information retrieval, recommender systems, natural language processing, language, cognition and computation, and other problems in dynamical systems. Young researchers presented their current work and had ample opportunity to interact with eminent professors and scholars in their area of research. The researchers also benefitted from the discussions that sparked new ideas in approaching their research problems. The authors of short/poster papers illustrated their work during a special session and obtained feedback from eminents.

A large number of eminent professors, well-known scholars, and young researchers participated in making MIKE 2013 a great success. We express our sincere thanks to the management of VHNSN College, particularly Thiru. P.C.S. Govindaraja Perumal, Secretary Correspondent, and Capt. Dr. P. Sundara Pandian, Principal, VHNSN College (Autonomous), Virudhunagar, who served as patron of this event. We were pleased to have Prof. Yutaka Maeda, Kansai University, Japan, Prof. Ramon Lopaz de Mantaras, Artificial Intelligence Research

Institute, Spain, Prof. Mandar Mitra, Indian Statistical Institute (ISI), Kolkata, India, Prof. Björn Gambäck, Norwegian University of Science and Technology, Norway, Dr. Philip O' Reilly, University College Cork, Ireland, and Dr. Santiago Ontanon, Drexel University, USA, serving as advisory chairs of MIKE 2013.

Several eminent scholars, including Prof. Ramon Lopaz de Mantaras, Artificial Intelligence Research Institute (IIIA), Spain, Prof. Pinar Ozturk, Norwegian University of Science and Technology(NTNU), Norway, Prof. Niloy Ganguly, Indian Institute of Technology, Kharagpur, India, and Prof. N. Subba Reddy, Gyeongsang National University, Korea, delivered invited talks on learning and knowledge exploration tasks in various interdisciplinary areas of science and technology. We also had an industrial talk delivered by Dr. Krishnaiah Jallu, Dy. Manager (R&D), BHEL Tiruchirappalli, India.

Dr. Prasenjit Majumdar, DAIICT, Gandhi Nagar, India, and Dr. Amitava Das, University of North Texas, USA, served as the workshop chairs. We thank the organizers of the Language, Cognition and Computation (LCC 2013) Workshop: Prof. Anupam Basu, Mr. Tirthankar Dasgupta, and Ms. Manjira Sinha, from the Communication Empowerment Laboratory, Indian Institute of Technology, and Dr. Sibansu Mukhopadhay, SNLTR, Kolkata, for their pains and efforts in orgranizing LCC 2013. We thank Dr. Amitava Das and Dr. Balamurali, A.R, Samsung Research India, for initiating steps to organize the First Workshop on Sentiment Analysis for Indian Languages (SAIL 2013). We also thank the invited speakers of these workshops: Dr. Kamal Kr. Choudhary, Indian Institute of Technology, Ropar, Rajastan, and Dr. Uttama Lahiri, Indian Institute of Technology, Gandhi Nagar, Gujarat.

We are very grateful to all our sponsors including FEXCO Financial Services, Ireland, University College Cork, Ireland, Elena Geo Systems, Government of India funding bodies, and especially the University Grants Commission, Department of Science and Technology, Defence Research and Development Organisation, All India Council for Technical Education, and the Tamil Nadu State Council for Science and Technology for their support of MIKE 2013.

We thank the Program Committee and the additional reviewers for their timely and thorough participation in the reviewing process. We appreciate the time and effort put in by the local organizers at the Research Center in Computer Science and the Department of Information Technology, VHNSN College, who dedicated their time to MIKE 2013. Finally, we acknowledge the support of EasyChair in the submission, review, and proceedings creation processes. We are very pleased to express our sincere thanks to Springer, especially Alfred Hofmann and Anna Kramer for their faith and support in publishing the proceedings of MIKE 2013.

December 2013

Rajendra Prasath
T. Kathirvalavakumar

Organization

Program Committee

Murugan A.	Dr. Ambedkar Government Arts College, India
Padmapriya A.	Alagappa University, India
Suriliandi A.	Manonmanium Sundaranar University, India
Juan A. Recio-García	University Complutense of Madrid, Spain
Agnar Aamodt	Norwegian University of Science and Technology, Norway
Rakesh Chandra Balabantaray	International Institute of Information Technology, India
Poorna Balakrishnan	SSS Jain College for Women, India
Biswanath Barik	TCS Innovation Lab, India
Pinaki Bhaskar	Jadavpur University, India
Pushpak Bhattacharyya	Indian Institute of Technology Bombay, India
Tamali Bhattacharyya	Indian Institute of Technology, Kharagpur, India
Joydeep Chandra	Indian Institute of Technology, Patna, India
Gladis Christopher	Presidency College, India
Guru D.S.	University of Mysore, India
Amitava Das	University of North Texas, USA
Dipankar Das	National Institute of Technology (NIT), India
Maunendra Desarkar	Samsung R&D Institute India, India
George Dharma Prakash Raj	Bharathidasan University, India
Aidan Duane	Waterford Institute of Technology, Ireland
Rohan Dutta	KIIT University, Bhubaneswar, India
Björn Gambäck	Norwegian University of Science and Technology, Norway
Debasis Ganguly	Dublin City University, Ireland
Niloy Ganguly	Indian Institute of Technology Kharagpur, India
Saptarshi Ghosh	Bengal Engineering and Science University India
Bharath Gopalaswamy	University of Illinois at Urbana-Champaign, USA
Sumit Goswami	DRDO, Government of India, India
Roshan Joymartis	Ngee Ann Polytechnic, Singapore

Muneeswaran K.	Mepco Schlenk Engineering College, India
Somasundaram K.	Gandhigram Rural Institute, India
Shikhar Kr. Sarma	Gauhati University, India
Ramon Lopez de Mantaras	IIIA - CSIC, Spain
Gethsiyal Augasta M.	Sarah Tucker College, India
Durairaj M.	Bharathidasan University, India
Sumathi M.	Sri Meenakshi Govt Arts College, India
Padma M.C.	PES College of Engineering, India
Yutaka Maeda	Kansai University, Japan
Prasenjit Majumdar	DAIICT, Gandhi Nagar, India
Aradhna Malik	Indian Institute of Technology, India
Mandar Mitra	Indian Statistical Institute, India
Hans Moen	Norwegian University of Science and Technology, Norway
Jian-Yun Nie	University of Montreal, Canada
Philip O'Reilly	University College Cork, Ireland
Santiago Ontañón	Drexel University, USA
Pinar Ozturk	Norwegian University of Science and Technology, Norway
Shanmugavadivu P.	Gandhigram Rural Institute, India
G.A.Vijayalakshmi Pai	PSG College of Technology, India
Partha Pakray	Norwegian University of Science and Technology, Norway
Rajarshi Pal	Institute for Development and Research in Banking Technology, India
Shyamosree Pal	Indian Statistical Institute, India
Sukomal Pal	Indian School of Mines, India
Chhabi Rani Panigrahi	Indian Institute of Technology Kharagpur, India
Ranjani Parthasarathi	Anna University - College of Engineering, India
Rajendra Prasath	University College Cork, Ireland
Pattabhi R.K. Rao	MIT Campus, Anna University, India
P.V. Rajkumar	University of Texas at San Antonio, USA
Muthu Ramakrishnan M.	Ngee Ann Polytechnic, Singapore
Subba Reddy	Gyeongsang National University, South Korea
Sudip Roy	Indian Institute of Technology Kharagpur, India
Arumuga Perumal S.	S.T. Hindu College, India
Sivakumar S.	CPA College, India

Achuthsankar S. Nair	University of Kerala, India
Sujankumar Saha	Birla Institute of Technology, India
Saurav Sahay	Intel Labs, USA
Sajal Sarkar	Indian Institute of Technology, India
Dipti Misra Sharma	International Institute of Information Technology, India
Vijay Sundar	MIT Campus, Anna University, India
Udayabaskaran S.	VelTech University, India
Tripti Swarnkar	Indian Institute of Technology, India
Jaisingh T.	Indian School of Mines, India
Kumaran T.	Government Arts College, India
Aravalluvan T.	Arumugam Pillai Seethai Ammal College, India
Kalaiselvi T.	Gandhigram Rural Institute, India
Kathirvalavakumar T.	VHNSN College (Autonomous), India
Meyyappan T.	Alagappa University, Karaikudi, India
Jaiprakash T. Lallchandani	The International Institute of Information Technology, India
Geetha T.V.	Anna University - CEG, India
Diana Trandabat	University "Al. I. Cuza" of Iasi, Romania
Yegnanarayanan V.	Velammal Engineering College, India
Shanmuga Velan	Indian Institute of Technology, India
Komathy Vimalraj	Easwari Engineering College, India
Anil Kumar Vuppala	International Institute of Information Technology, India
Xiaolong Wu	California State University, USA

Additional Reviewers

A., Padmapriya
A., Suruliandi
Ahmad, Riyaz
Airola, Antti
Aradhya Vn., Manjunath
Augasta M., Gethsiyal
B.S., Harish
Bag, Soumen
Banerjee, Somnath
Bera, Sahadev
Bhaskar, Pinaki
Bhattacharya, Tamali
Chowdhury, Manish
Christopher, Gladis
Cordier, Amélie
Das, Amitava
Das, Gautam K.
Dogra, Debi
Ghosh, Saptarshi
Goswami, Sumit
Høverstad, Boye
J., Krishnaiah
J.Woodham, Robert
K., Somasundaram
K., Vinay
Kar, Pushpendu
Kisku, Bapi
Kohli, Shruti
Kr. Sarma, Shikhar
Lohar, Pintu
M., Durairaj
M., Sumathi

M., Durairaj
Maeda, Yutaka
Ontanon, Santiago
Pakray, Partha
Pal, Dipasree
Pal, Santanu
Pal, Shyamosree
Pandey, Basant Kumar
Pandey, Kumar
Pati, Bibudhendu
Patra, Braja Gopal
Pichai, Shanmugavadivu
Prakash Raj E., George Dharma
Roy, Rahul
S., Arumuga Perumal
Sang Woo, Kim
Shantharamu, Manjunath
Sharma, Dipti
Sharma, Dipti Misra
Sharma, Himanshu
Swaminathan, Udayabaskaran
Swarnkar, Tripti
T., Jaisingh
T., Kathirvalavakumar
T., Kumaran
T., Meyyappan
Torra, Vicenc
Vuppala, Anil Kumar
Wu, Huayu
Yegnanarayanan, Venkataraman

Table of Contents

A Feature Selection Method Using Hierarchical Clustering

Cheong Hee Park

Dept. of Computer Science and Engineering
Chungnam National University
Yuseong-gu, 305-764, Korea

Abstract. Feature selection refers to a problem to select a subset of features which are most optimal for intended tasks. As one of well-known feature selection methods, clustering features into several groups and picking one feature from each group have been used for unsupervised feature selection. Since the purpose of clustering in feature selection is to select a feature from each group, the quality of the feature to be selected should be considered in the clustering process. In this paper, we propose a feature selection method using hierarchical clustering. A new similarity measure between two feature groups is defined by directly using the representative feature in each group. Experimental results show that our method can select good features even for supervised learning.

Keywords: Feature selection, Hierarchical clustering, Ward method.

1 Introduction

Feature selection refers to a problem to select a subset of features which are most optimal for intended tasks such as classification, clustering or regression [1]. Reducing the dimensionality of data by feature selection can give some intuition about data property and also facilitate data mining processes afterwards. Feature selection methods are usually categorized to two groups: filter methods and wrapper methods. Filter methods evaluate the goodness of each feature according to some criterion and select features ranked with high scores. Wrapper methods utilize a learning algorithm in a selection process: a feature subset is searched which enhances the performance by the learning algorithm.

Feature selection methods can also be divided into two groups, supervised and unsupervised methods, according as class labels are used in a training stage or not. When the intended task is classification, the training data used in feature selection is labeled one and optimal features are the ones that discriminate classes well. Clustering features into several groups and picking one feature from each group lead to unsupervised feature selection in that class labels are not needed during the clustering process. It only requires a similarity measure between features. By choosing one feature from each group which similar feature patterns belong to, redundancy among features can be reduced.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 1–6, 2013.
© Springer International Publishing Switzerland 2013

In [2–4], hierarchical clustering was used for feature selection where the most similar two groups are merged and it is repeated until all features belong to one group. However, traditional methods which compute the similarity between two groups using complete linkage, single linkage, or ward method can not produce optimized results for feature selection, since they do not pay attention to the feature to be selected in each group in the final stage. In this paper, we propose a feature selection method using hierarchical clustering. The proposed method computes the similarity between two feature groups by directly using the representative feature in each group. The representative feature in each group is defined as the one which has the highest average similarity to other features in the group.

The rest of the paper is composed as follows. In Section 2, related works are reviewed. In Section 3, we propose a new method for feature selection using hierarchical clustering. Experimental results demonstrate that the proposed method is superior to other compared methods in Section 4. Discussions follow in Section 5.

2 Related Works

When no class labels are available, feature selection methods are called as unsupervised feature selection. Laplacian score (LS) is a well-known unsupervised feature selection method [5]. LS evaluates the features according to their locality preserving power. For the r-th feature $f_r = [f_{r1}, f_{r2}, \cdots, f_{rm}]$ and the weight matrix between data samples $S = [S_{ij}]_{1 \le i,j \le m}$, the locality preserving power of the feature f_r is considered to be high, when the value

$$\sum_{ij} (f_{ri} - f_{rj})^2 S_{ij} \tag{1}$$

is minimized. However, when the training data with class labels is given, feature selection reflecting class distribution should be better for classification than unsupervised feature selection. Fisher score (FS) is a representative filter method which works independently with any classifier [6]. Features which maximize the between-class variance and minimize the within-class variance get high fisher scores. Fisher score for the r-th feature f_r is computed as

$$\frac{\sum_{i=1}^{c} n_i (\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2}, \tag{2}$$

where μ is the mean of the whole data set and μ_i and σ_i^2 are the mean and variance of the class i corresponding to the feature f_r. n_i is the number of the elements in class i. Since the LS or FS scores are computed independently for each individual feature, they are simple and fast to implement. However, independent processing of features makes it difficult to eliminate redundancy among features.

Hierarchical clustering produces grouped feature sets and it is easy to perform feature selection by choosing one feature from each feature group. The

most well-known similarity measure between random variables corresponding to feature vectors is correlation. In [2], starting with groups composed of individual feature vector, at each pass the two groups with the highest correlation are merged. Each feature group is represented by the sum or centroid of the feature vectors in the group. The authors recognized that their clustering technique, denoted as SWC (step-wise clustering), is same as clustering by ward method [7] where the objective function is the pairwise correlation coefficient between group centroids. Hierarchical clustering by Ward method was also used in [4], where the distance matrix using Barthelemy-Montjardet distances was designed. However, the method is limited to nominal attributes in practice. In [3], hierarchical clustering by complete linkage was applied for feature selection in spectral data by merging the two most similar consecutive feature groups with the correlation measure.

Since the final purpose of clustering is to select a feature from each group, the quality of the feature to be selected should be considered at the stages to choose two groups to merge. We propose an objective function to utilize the representative feature vector which has the highest average similarity to other feature vectors in the group.

3 Feature Clustering by the Modified Ward Method

Ward method merges two groups G_r and G_s among all the pairs of groups, which minimize

$$D(G_r, G_s) = \sum_{x \in G_r \cup G_s} \|x - \bar{x}\|^2 - \Big(\sum_{x \in G_r} \|x - \bar{x_r}\|^2 + \sum_{x \in G_s} \|x - \bar{x_s}\|^2 \Big). \quad (3)$$

Here, $\bar{x_r}$ and $\bar{x_s}$ are the centroids of G_r and G_s respectively, and $\bar{x}$ is the centroid after the merge of G_r and G_s. As in Eq. (3), the centroid has been used as the one to represent the set of data points in many algorithms such as Linear Discriminant Analysis and K-means algorithm[6]. The centroid is not a real data point in the set, but it minimizes the sum of the squared distances to all other data points in the set. When we need a real data point, the medoid can be used instead of the centroid. The medoid is the representative point of the set whose average dissimilarity to all the points in the set is minimal. We utilize the concept of medoid in defining an objective function for hierarchical clustering and also in choosing the representative feature in each feature group.

Let $s(\cdot, \cdot)$ be similarity measure between two feature vectors. Correlation or absolute value of correlation can be used as a similarity measure. Each feature group is represented with the feature vector whose similarity sum to other feature vectors in the group is maximum. In other words, the representative feature vector in the group G_r is defined as

$$\hat{f} = argmax_{f \in G_r} \sum_{g \in G_r,\ g \neq f} s(f, g), \quad (4)$$

and the power of $\hat{f}$ is measured as $\sum_{g \in G_r,\ g \neq \hat{f}} s(\hat{f}, g)$. Now, we define the similarity between two feature groups G_r and G_s as follows.

$$s(G_r, G_s) = \max_{f \in G_r} \sum_{g \in G_r,\ g \neq f} s(f, g) \tag{5}$$
$$+ \max_{f \in G_s} \sum_{g \in G_s,\ g \neq f} s(f, g) - \max_{f \in G_r \cup G_s} \sum_{g \in G_r \cup G_s,\ g \neq f} s(f, g).$$

The first and second terms in Eq. (5) compute the sum of similarities between any feature vector and all other feature vectors within each group and find the greatest value. The third term performs the same calculation for the case after the merge of two groups. Hence, the objective function in Eq. (5) computes the reduction in the power of the representative feature vectors by the merge of two groups. In the hierarchical clustering process, two feature groups which minimize the objective function by Eq. (5) are merged.

Feature selection using Hierarchical clustering can be distinguished by two steps of defining an objective function for hierarchical clustering and choosing the representative feature in each feature group. By applying Eq. (4) and (5) in hierarchical clustering process, we define two feature selection algorithms, modSWC1 and modSWC2:

- SWC [2] :
  - Repeatedly merge the two groups with the highest correlation between the centroids of the feature vectors in the group until only one group remains.
  - In the dendrogram by the hierarchical clustering, find the point where k groups are built. From each group, select one feature vector which has the highest correlation with the centroid of the group.
- modified version 1 of SWC (modSWC1) :
  - Repeatedly merge the two groups with the smallest value by Eq. (5) until only one group remains.
  - In the dendrogram by the hierarchical clustering, find the point where k groups are built. From each group, select one feature vector by Eq. (4), the sum of whose similarities to other feature vectors in the group is maximum.
- modified version 2 of SWC (modSWC2) :
  - Repeatedly merge the two groups with the highest correlation between the centroids of the feature vectors in the group until only one group remains.
  - In the dendrogram by the hierarchical clustering, find the point where k groups are built. From each group, select one feature vector by Eq. (4), the sum of whose similarities to other feature vectors in the group is maximum.

The modified version 1 of SWC is the proposed method utilizing the Eqs. (4) and (5), while the modified version 2 of SWC is the method which uses only Eq. (4).

Table 1. Data description

Data set	samples	classes	features
Spam	4601	2	57
Optdigits	5620	10	64
Isolet	7798	26	617

4 Experimental Results

In order to test the performances of three feature selection methods using hierarchical clustering, we used data sets downloaded from UCI machine learning repository. Three data sets used in the experiments are described in Table 1. Three methods, SWC, modSWC1 and modSWC2, are compared with the feature selection using Fisher score. Classification performance of each method was measured

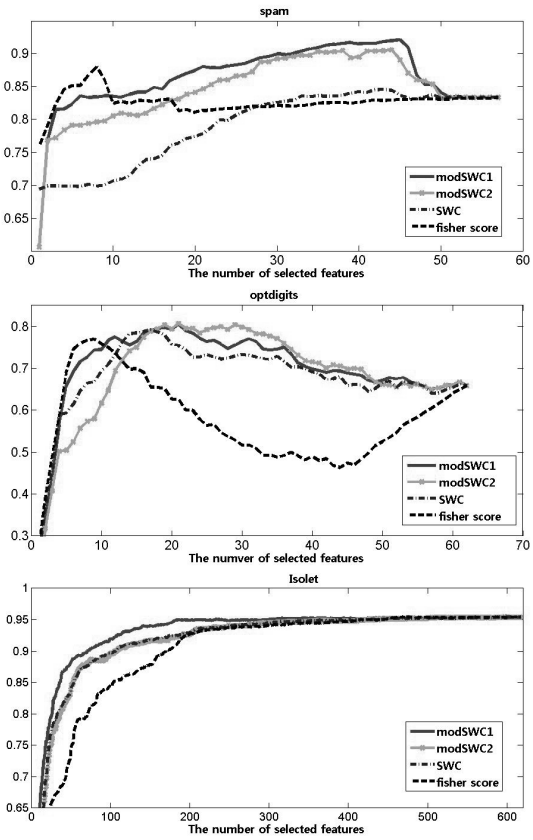


Fig. 1. Comparison of prediction accuracies

using the SVM classifier by 5-cross validation. Each data set was split to the training set and test set at each iteration of 5-cross validation. Hierarchical clustering was performed on the training set and the given number of features were selected. A SVM classifier with a Gaussian kernel was modeled on the training set only with the selected features and the prediction accuracy was measured on the test set. A software libSVM (`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`) was used for SVM with default parameter settings. For feature selection by Fisher score, the training set was used to compute fisher scores and the SVM classifier was modeled with the selected features.

The average accuracies by cross validation are displayed in Figure 1. For all three data sets, the algorithms, modSWC1 and modSWC2, are better than the original SWC and the method by Fisher score. Note that even though modSWC1 and modSWC2 do not take advantages of class labels, they outperform the method to use Fisher score.

5 Discussions

In this paper, we proposed feature selection methods using Hierarchical clustering. Since the final purpose of clustering is to select a feature from each group, the quality of the feature to be selected should be considered at the stage to choose two groups to merge. We proposed an objective function to utilize the representative feature vector which has the highest average similarity to other feature vectors in the group. Experimental results show that our method can select good features for supervised learning.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2011-0007779).

References

1. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. 19, 153–157 (2000)
2. King, B.: Step-wise clustering procedures. Journal of the American Statistical Association 62(317), 86–101 (1967)
3. Krier, C., Francois, D., Rossi, F., Verleysen, M.: Feature clustering and mutual information for the selection of variables in spectral data. In: Proceedings of European Symposium on Artificial Neural Networks, Bruges, Belgium (2007)
4. Butterworth, R., Piatetsky-Shapiro, G.: On feature selection through clustering. In: Proceedings of the Fifth IEEE International Conference on Data Mining (2005)
5. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada (2005)
6. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley Interscience, New York (2001)
7. Ward Jr., J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)

Rank Aggregation for Filter Feature Selection in Credit Scoring

Waad Bouaguel[1], Ghazi Bel Mufti[2], and Mohamed Limam[1,3]

[1] LARODEC, ISG, University of Tunis,Tunisia
[2] LARIME, ESSEC, University of Tunis,Tunisia
[3] Dhofar University, Oman
bouaguelwaad@mailpost.tn,belmufti@yahoo.com,
mohamed.limam@isg.rnu.tn

Abstract. The credit industry is a fast growing field, credit institutions collect data about credit customer and use them to build credit model. The collected information may be full of unwanted and redundant features which may speed down the learning process, so, effective feature selection methods are needed for credit dataset. In general, Filter feature selection methods outperform other feature selection techniques because they are effective and computationally fast. Choosing the appropriate filtering method from the wide variety of classical filtering methods proposed in the literature is a crucial issue in machine learning. So, we propose a feature selection fusion model that fuses the results obtained by different filter feature selection methods via aggregation techniques. Evaluations on four credit datasets show that the fusion model achieves good results.

Keywords: Feature selection, filter, aggregation, error curve.

1 Introduction

Many empirical studies show that manipulating few variables in credit scoring leads certainly to more reliable and better understandable models without irrelevant, redundant and noisy data [1]. The more the number of features grows the more computation is required and model accuracy and scoring interpretation are reduced [2]. To overcome these problems we perform a feature selection on the original features set.

In feature selection process we choose an appropriate feature subset that contains the most relevant features. A variety of techniques to select the best subset of features have been proposed. Three main classes of feature selection are identified in the literature as stated by [3,4]: filter, wrapper and hybrid feature selection methods. A filter technique is a pre-selection process which is independent of the later applied classification algorithm. Filters can be exceptionally effective because they need to be performed only once without any search involved. A wrapper technique on the other hand uses specific classifier and exploits resulting classification performance to select features. This kind of methods use search

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 7–15, 2013.
© Springer International Publishing Switzerland 2013

techniques to pick subsets of variables and evaluate their importance based on the estimated classification accuracy [4]. The hybrid approach uses both filtering and wrapping methods for improving the performance of the feature selection.

According to [5] filters methods outperforms other feature selection methods in many cases. There are a variety of classical filter methods in previous literature [1,6]. Given the variety of techniques, the question is how to choose the best one for a specific feature selection task? [5] call this problem a selection trouble. Hence, we propose to investigate on a new fusion framework. In this paper we focus on combining different filtering criteria into a new result in order to obtain a better rank list, by using a aggregation rules. This paper is organized as follows. Section 2 reviews filter feature selection methods and features aggregation. Section 3 gives experimental results on four datasets and in Section 4 conclusions are drawn.

2 Selection Trouble

2.1 Filter Feature Selection Method

The basic idea of filter methods is to select the best features according to some prior knowledge. Filter feature selection methods can be grouped into two categories, i.e. feature weighting methods and subset search methods. This categorization is based on whether they evaluate the relevance of features separately or through feature subsets. In feature weighting methods, weights are assigned to each feature independently and then the features are ranked based on their relevance to the target variable. Relief is a famous algorithm that study features relevance [7]. This method uses the Euclidean distance to select a sample composed of a random instance and the two nearest instances of the same and opposite classes. Then a routine is used to update the feature weight vector for every sample triplet and determines the average feature weight vector relevance. Then, features with average weights over a given threshold are selected.

Subset search methods explore all possible feature subsets using a particular evaluation measure. The best possible subset is selected when the search stops. According to [8], consistency and correlation [9,10] are the best evaluation measures that decrease efficiently irrelevance and redundancy. A Consistency measure evaluates the distance of a feature subset from the consistent class label. Consistency is established when a data set with the selected features alone is consistent. That is, no two instances may have the same feature values if they have a different class label [10]. A correlation measure is applied between two features as a goodness measure. That is a feature is considered as good if it is highly correlated to the class and uncorrelated with any other features. [8] recommended two main approaches to measure correlation, the first one is based on classical linear correlation between to random variables and the second one is based on information theory.

Numerous correlation coefficients can be used under to first approach but the most common is the Pearson correlation coefficient (PCC). PCC is a simple measure that has been shown to be effective in a wide variety of feature selection

methods ([4]). Formally, the PCC for two continuous random variables x_i and x_j is defined as :

$$PCC = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}}, \tag{1}$$

where where cov is the covariance of variables and var is the variance of each variable. Simple correlation measure in general measures the linear relationship between two random variables, which may be not suitable in some cases. The second approach based on information theory measures how much knowledge two variables carry about each other. Mutual information (MI) is a well known information theory measure that captures nonlinear dependencies between variables. Formally, the mutual information of two continuous random variables x_i and x_j is defined as:

$$I(x_i, x_j) = \int\int p(x_i, x_j) log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} dx_i dx_j, \tag{2}$$

where $p(x_i, x_j)$ is the joint probability density function, and $p(x_i)$ and $p(x_j)$ are the marginal probability density functions.

The majority of above cited features selection methods select the k top ranked features. In general, filter criteria are used independently. That is, one feature selection method is employed and performance is measured according to the selected features. The question is then which method will be the most appropriate to our study. Rather than to study what each single criterion can offer, we can employ these methods in combination.

3 Ensemble Features Methods

3.1 Filter Feature Selection Aggregation

Two effective modes to fuse a set of filtering feature selection methods are proposed in the literature [5]. In the first mode, the final outputs of each single filter method are combined into a one single result. The second fusion mode, combine the different filtering criteria of each filter method in order to find a new measure that select the best feature subset. In general, when the second mode is used, we not only need some prior knowledge about the data but also a familiarity with the criteria to be combined and good mathematical skills, therefore the first mode is choose over the second, because it is the simplest one and because it does not require additional configuration. In order to implement the chosen fusion mode, aggregation techniques can be used.

The main thought behind using ensemble feature aggregation is to obtain a list of significant and jointly selected set of features that can be used during the classification process. We try in this context to capture features which may provide essential factors during the prediction of the credit-worthiness by removing the redundant ones. Typically ensemble feature aggregation reduce the

biases caused by individual feature algorithms while providing higher accuracy, sensitivity, and specificity, which are often not achievable with individual feature selection techniques or while not using any feature selection techniques at all.

In general, when we deal with aggregating feature rankings, there are two issues to consider. The first one is which base feature rankings to aggregate. There are different ways to generate the base feature rankings:

- using the same dataset but by different filter methods.
- using different datasets but the same filtering method.
- using different subsamples of the same dataset and the same ranking method.

The second issue concerns the type of aggregation function to use. Ensemble selection consists of multiple runs of feature ranking which are then combined into a single ranking for each feature. One of the most critical decisions when performing ensemble feature selection is deciding on which aggregation technique to use for combining the resulting ranked feature lists from the multiple runs of feature ranking into a single decision for each feature.

For the first issue we decide to use the same dataset with different filter methods. The three previously discussed feature selection criteria namely relief, PCC and MI are then considered. For the second issue many functions are available in the literature, like taking the mean or median of the ranks. This paper is an in-depth comparison between two aggregation techniques: Majority Vote and Mean Aggregation.

Majority vote is a common classifier combination method, particularly used in classifier ensembles when the class labels of the classifiers are crisp [11]. In general, majority voting is a simple method that does not require any parameters to be trained or any additional information for the later results [3]. We propose to use majority voting to feature selection in order to fuse an ensemble of filter methods. This method use voting for selecting the features with the major amount of votes. In this case the input is a set of ranking lists generated by several feature selection techniques, and which are sorted in descending order according to their corresponding votes, from the most significant feature to the least one. The output is a single list of features corresponding to the most discriminating features.

Mean Aggregation consists of taking the average rank across all of the ranked feature lists and using that mean value to determine the final rank of the feature. Mean aggregation technique is practical and easy to implement which make it frequently used for ensemble feature selection [12].

3.2 Error Curve

Once the selection trouble is resolved and a consensus list of mutual features is obtained, we come across the issue of choosing the appropriate number of features to retain. In fact a list of sorted features doesn't provide us with the optimal features subset. In general a predefined small number of features is retained from the consensus list for constructing the final model. If the number

of used features is relatively small or big, then the final classification results may be a degraded.

In this section, we approach the problem of choosing the appropriate number of features by following the idea that the precision of the feature rank is related to predictive accuracy. In fact aggregation would put on top of a list a feature that is most important, and at the bottom a feature that is least important relatively to the target concept. All the other features would be in-between, ordered by decreasing importance. By following this intuition, we choose the number of the most pertinent features by performing a stepwise feature subset evaluation, with which we generate a so-called error curve. We rely on the process of generating the error curve (Figure 1). We begin with the obtained ranked list in Section 3, we then construct the credit model with only the top-ranked feature and we then add to this feature the second ranked feature . This process is continued iteratively until a bottom ranked feature is added yielding to decrease in the general accuracy. The points of the error curve are each of the n estimated errors and the point where the error curve decrease is considered as the selection boundary for the appropriate number of features.

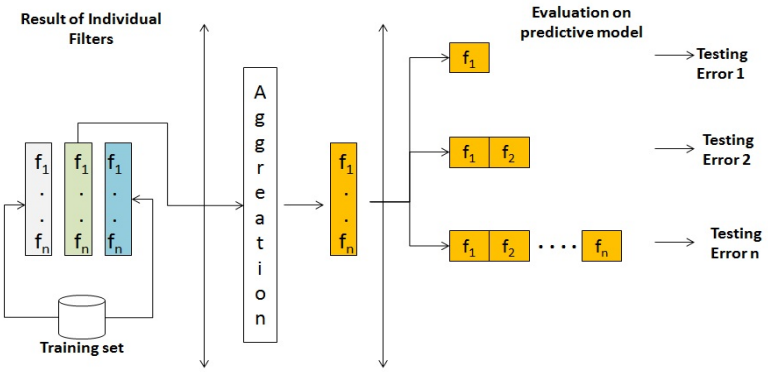


Fig. 1. Ensemble feature selection

4 Empirical Study

Four real-world datasets with detailed input attributes description are selected to study the performance of the proposed approach: two datasets from the UCI repository of machine learning databases (i.e. Australian and German credit datasets) and a dataset from a Tunisian bank and the HMEQ dataset.

– The Australian dataset present an interesting mixture of attributes: 6 continuous, 7 nominal and a target attribute with few missing values. This dataset is composed of 690 instances where 306 ones are creditworthy and 383 are not. All attribute names and values have been changed to meaningless symbols for confidentiality.

- The German credit dataset covers a sample of 1000 credit consumers where 700 instances are creditworthy and 300 are not. For each applicant, 21 numeric input variables are available .i.e. 7 numerical, 13 categorical and a target attribute.
- The HMEQ dataset covers a sample of 5960 instances describing recent home equity loans where 4771 instances are creditworthy and 1189 are not. The target is a binary variable that indicates if an applicant is eventually defaulted. For each applicant, 12 input variables were recorded where 10 are continuous features, 1 is binary and 1 is nominal.
- The Tunisian dataset covers a sample of 2970 instances of credit consumers where 2523 instances are creditworthy and 446 are not. Each credit applicant is described by a binary target variable and a set of 22 input variables were 11 features are numerical and 11 are categorical. Table 1 displays the characteristics of the datasets that have been used for evaluation.

Table 1. Results summary for the Australian dataset

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.923	0.923	0.923	0.923
PCC	0.924	0.926	0.924	0.926
MI	0.944	0.919	0.944	0.929
Majority	0.946	0.926	0.946	0.934
Mean	0.934	0.927	0.934	0.931
	NB			
Relief	0.88	0.941	0.88	0.909
PCC	0.935	0.918	0.935	0.927
MI	0.944	0.903	0.944	0.923
Majority	0.945	0.948	0.932	0.929
Mean	0.943	0.923	0.943	0.928
	SVM			
Relief	0.880	0.941	0.88	0.909
PCC	0.880	0.931	0.86	0.905
MI	0.890	0.910	0.908	0.890
Majority	0.908	0.931	0.908	0.910
Mean	0.908	0.931	0.890	0.910

In general mutual information computation requires estimating density functions for continuous variables. For simplicity, each variable is discretized. Then, we split the datasets into a training sample and a test sample, where the first deals with the new feature selection approach and the diverse classification models and the second one checks the reliability of the constructed models in the learning step. The experimental study compares the performance of the fusion approach with the individual filter methods.

The performance of our system is evaluated using the True positive (TP) and False positive (FP) rates and the standard Information retrieval (IR) performance measures: Precision, Recall and F-measure metrics. Results summarized

Table 2. Results summary for the German dataset

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.511	0.692	0.511	0.588
PCC	0.5	0.721	0.500	0.591
MI	0.580	0.750	0.580	0.654
Majority	0.578	0.781	0.586	0.658
Mean	0.578	0.781	0.586	0.656
	NB			
Relief	0.5	0.638	0.5	0.561
PCC	0.477	0.737	0.477	0.579
MI	0.523	0.742	0.523	0.613
Majority	0.556	0.716	0.545	0.619
Mean	0.542	0.750	0.542	0.612
	SVM			
Relief	0.489	0.694	0.489	0.573
PCC	0.489	0.705	0.489	0.577
MI	0.545	0.738	0.545	0.627
Majority	0.557	0.766	0.557	0.645
Mean	0.552	0.766	0.552	0.627

Table 3. Results summary for the HMEQ dataset

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.836	0.819	0.836	0.81
PCC	0.974	0.838	0.974	0.901
MI	0.836	0.819	0.836	0.81
Majority	0.968	0.853	0.976	0.912
Mean	0.966	0.850	0.966	0.904
	NB			
Relief	0.8	0.747	0.8	0.736
PCC	0.832	0.818	0.832	0.798
MI	0.831	0.814	0.831	0.801
Majority	0.97	0.843	0.97	0.902
Mean	0.981	0.821	0.981	0.887
	SVM			
Relief	0.807	0.845	0.807	0.728
PCC	0.828	0.822	0.828	0.784
MI	0.828	0.822	0.828	0.784
Majority	0.989	0.835	0.989	0.905
Mean	0.987	0.830	0.987	0.902

Table 4. Results summary for the Tunisian dataset

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.848	0.827	0.847	0.830
PCC	0.850	0.833	0.850	0.832
MI	0.852	0.822	0.852	0.826
Majority	0.985	0.866	0.985	0.921
Mean	0.964	0.875	0.964	0.917
	NB			
Relief	0.888	0.876	0.888	0.882
PCC	0.880	0.876	0.880	0.879
MI	0.883	0.885	0.883	0.884
Majority	0.981	0.866	0.981	0.920
Mean	0.960	0.860	0.962	0.913
	SVM			
Relief	0.85	0.722	0.85	0.781
PCC	0.847	0.769	0.847	0.784
MI	0.994	0.851	0.994	0.917
Majority	0.998	0.849	0.999	0.930
Mean	0.993	0.840	0.994	0.927

in each Table 1 and Table 2 represent the performance of each feature selection technique for three different classification techniques: Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machine (SVM).

Tables 1-4 summarize the performances achieved by LR, NB, and SVM algorithms using 3 individual filters namely relief, PCC, MI and their majority vote aggregation and mean aggregation. A more detailed picture of the achieved results shows that in most cases, aggregation approaches usually outperform single filters.

5 Conclusion

In this study, we investigate on merging filter feature selection methods within a credit scoring framework. Our work was conducted on two parts. First, we conducted a preliminary study on two rank aggregation approaches, namely majority voting and mean aggregation. Second we investigated on choosing the right number of features from the final ranked list, we evaluated the ranking by performing a stepwise feature subset evaluation, resulting on an error curve. Results show that there is a generally beneficial effect of aggregating feature rankings as compared to the ones produced by single methods. In fact the fusion performance is either superior to or at least as close as either of filter methods. In additional to this work, selecting the right number of features is a challenge, however to select the appropriate number of feature from a ranking list is still an open problem to be studied in the future. In our further work we plan to go beyond the visual inspection of the error curves. The first step would be to use the area under the error curve as a metric to evaluate the quality of the curves.

References

1. Wang, C.M., Huang, W.F.: Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. Expert Syst. Appl. 36, 5900–5908 (2009)
2. Howley, T., Madden, M.G., O'Connell, M.L., Ryder, A.G.: The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowl.-Based Syst. 19, 363–370 (2006)
3. Guldogan, E., Gabbouj, M.: Feature selection for content-based image retrieval. Signal, Image and Video Processing, 241–250 (2008)
4. Rodriguez, I., Huerta, R., Elkan, C., Cruz, C.S.: Quadratic Programming Feature Selection. Journal of Machine Learning Research 11, 1491–1516 (2010)
5. Wu, O., Zuo, H., Zhu, M., Hu, W., Gao, J., Wang, H.: Rank aggregation based text feature selection. In: Web Intelligence, pp. 165–172 (2009)
6. Bouaguel, W., Bel Mufti, G.: An improvement direction for filter selection techniques using information theory measures and quadratic optimization. International Journal of Advanced Research in Artificial Intelligence 1, 7–11 (2012)
7. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of the Ninth International Workshop on Machine Learning, pp. 249–256. Morgan Kaufmann Publishers Inc., San Francisco (1992)
8. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML, pp. 856–863 (2003)
9. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: ICML, pp. 359–366 (2000)
10. Arauzo-Azofra, A., Benitez, J.M., Castro, J.L.: Consistency measures for feature selection. J. Intell. Inf. Syst. 30(3), 273–292 (2008)
11. Kuncheva, L.I., Bezdek, J.C., Duin, P.W.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 34, 299–314 (2001)
12. Wald, R., Khoshgoftaar, T.M., Dittman, D.J.: Mean aggregation versus robust rank aggregation for ensemble gene selection. In: ICMLA (1), pp. 63–69 (2012)

Hybrid Approach for Palmprint Recognition Using Compound Features

N.L. Manasa, A. Govardhan, and Ch. Satyanarayana

Jawaharlal Nehru Technological University, Andhra Pradesh, India

Abstract. As patterns in a palmprint have abundance of invariance, the inter-class and intra-class variability of these features makes it difficult for just one set of features to capture this variability. This inspires us to propose a hybrid feature extraction and fusion approach for palmprint recognition based on texture information available in the palm. Scale, shift and rotation (Affine) invariance, good directional sensitivity properties of Dual-tree Complex Wavelets makes it a choice to capture texture features at global level. Local Binary Pattern on the other hand being gray-scale and rotation invariant, captures local fine textures effectively. These local features are sensitive to position and orientation of the palm image. Canonical Correlation Analysis is used to combine the features at the descriptor level which ensures that the information captured from both the features are maximally correlated and eliminate the redundant information giving a more compact representation. Experimental results demonstrate an accuracy of 97.2% at an EER of 3.2% on CASIA palmprint database.

1 Introduction

Human vision system uses both global and local features to recognize the object of interest, and hybrid approaches are thus expected to be promising for palmprint recognition [21]. While ridge like patterns, valleys, minutae points and pores can be extracted only from a high-resolution image, with at least 400 dpi (dots per inch), features like principal lines and wrinkles can be extracted from a low-resolution image, with less than 100 dpi. Though ridge based authentication systems exist for latent palmprint recognition [7], the time consumed in image acquisition and processing restrict them from being used extensively in civil applications.

Creases/Palm lines, the evident structural features on the palm are formed several months after conception. Principal lines which are a result of genetic effects, though are significant, they alone cannot represent the uniqueness of a palmprint as twins and any two people can have similar principal lines as depicted in Fig 2. Palm wrinkle based authentication systems are counter-productive as the constancy of wrinkles can be compromised as they have chances of vanishing/diminishing over time or with extensive physical work with hands. Also few wrinkles are only stable for several months or years after conception. Palm line based approaches [9][10] which rely on derivatives of Gaussian and extraction of

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 16–27, 2013.
© Springer International Publishing Switzerland 2013

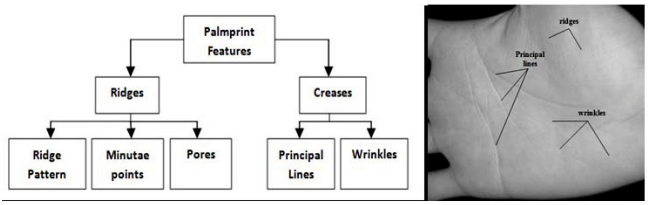


Fig. 1. Different kinds of features on a palmprint

width and location information of palm lines respectively were proposed which suffer from the non-permanence of the chosen features. Successful methods like Palmcode[5], Fusion code[18] and Competitive code[19] which rely on palmlines and their orientations, though attain high accuracy and low error rates, are subject to instability and are not self-sustained for the above reasons. Structural similarities among Palmcodes from palms can be observed from [2].

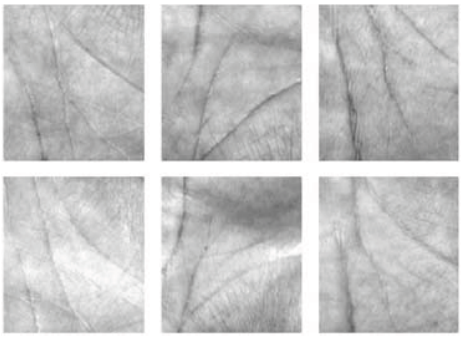


Fig. 2. Images of different individuals with similar principal lines. Courtesy [1].

Along with being permanent, a feature for human authentication should exhibit exclusiveness and idiosyncrasy. It should prove large variance between persons and small variance between samples of the same palm. Another feature that can be extracted from low-resolution palmprint images is Texture. Low-resolution palmprint images can also be considered as texture images. Texture is one of the clearly observable, permanent, distinguishable features on the human palm. This peculiarity inspires us to pursue a reliably working palmprint authentication system based on texture description. Palmprint samples from different individuals having distinctive texture features can be seen from Fig 3.

Texture on the palm can be clearly described by both, features at the local level and at the global level. Global palmprint features possess the following characteristics: (a) insensitive to noise; (b) insensitive to shift changes; (c) easy to compute; and 4) have high convergence within the group and good dispersion between groups [4]. Global features can lower the effect raised by local noises thus supplementing each other. Motivation for the proposed approach is as follows,

1. Although palmprint patterns are diverse among individuals, it is very difficult to distinguish solely based on global texture features as some of these patterns are so similar at a coarse level.
2. Most of the widely popular coding-based methods like palmcode [5] neglect the multiscale characteristic of palm lines [3] and construct authentication systems based on structural similarity.
3. In a peg-free and unconstrained acquisition environment, translation and rotation variations are inevitable. Image description at a local level handles such interferences reasonably [6].

A hand-based hierarchical authentication system [11] was developed to ensure fast matching which uses hand-geometry features at the coarse-level which are unstable in unconstrained environments. [8] demonstrated that the way to improve performance is intra-modal combination of texture-based, line-based and appearance-based features in the palm. They used various score-level and decision-level fusion techniques and claimed the superior performance of product of sum rule which still had higher error rates. [4] designed a hierarchical palmprint recognition approach and thus inferred that local features perform better than hierarchical approach when false acceptance rate is more than 5%.

Hence, all these factors inspired us to propose a hybrid approach which performs a feature-level fusion of local and global texture features of the palm. The above methods fuse the features at the image level where as the proposed method maximizes the correlation between the feature sets at feature level. Although a palmprint authentication system consists of several stages viz., Image Acquisition, Preprocessing/Region-of-interest (ROI) extraction, Feature extraction and Matching, the first two stages are out of the scope of this paper. Image Acquisition can be referred from [12] and ROI extraction proceeds similarly as stated in [1]. Distinctive from the literature, we propose the following hybrid approach to

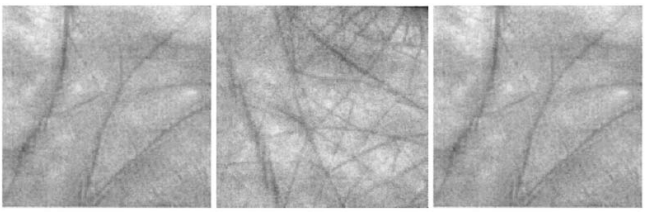


Fig. 3. Samples of different palmprint patterns with distinctive texture features (left) Strong principal lines. (middle) Less wrinkles. (right) Strong wrinkles. Courtesy [1]

overcome the above mentioned challenges and drawbacks. A feature descriptor fusion method based on Canonical Correlation Analysis is used to combine two features, a global feature set and a local feature descriptor: Dual-tree complex wavelets [DTCW][16] and Local Binary Patterns (LBP)[15] respectively. Canonical Correlation Analysis (CCA) is one of the important statistical multi-data processing methods which deals with the mutual relationships between two random vectors. This allows us to capture not only the crease features of the palm

but also captures ridge features, palm-line orientation information and magnitude features as all these can be perceived as local texture [3].

As patterns in the palmprint have abundance of invariance, the inter-class and intra-class variability of these features makes it difficult for just one set of features to capture this variability. The Global features are insensitive to affine transformations, noise, and captures large between-class variance and small within-class variance while Local features capture significant within-class variance. DTCW captures the global information ensuring scale-invariance and shift-invariance which helps in discriminating between locally similar regions. All these properties along with its good directional selectivity in 2D ensure favorable recognition of similar patterns [13]. DTCW features compensate the error in localizing the palm region as they are invariant to the rotation and the inexact localization. The LBP on the other hand captures local fine textures effectively, they are also sensitive to position and orientation of the palm image. It is a powerful texture descriptor that is gray-scale and rotation invariant [14]. A chance of recognition rate being compromised in the case of very large databases is high with different palmprint images having similar global features. Hence it is important to use local texture features in combination with global texture features for accurate recognition.

The proposed method relies on coarse ROI localization and extracts both the feature descriptors. Canonical correlation analysis is used to combine the features at the descriptor level which ensures that the information captured from both the features are maximally correlated and eliminate the redundant information giving a more compact representation.

2 Global and Local Feature Extraction

2.1 Dual-Tree Complex Wavelet Features

Discrete Wavelet Transforms based methods have been successfully applied to a variety of problems like denoising, edge detection, registration, fusion etc., Discrete wavelet transforms have 4 basic problems such as Oscillation, Shift variance, Aliasing and Lack of directionality. By using complex valued basis functions instead of real basis functions these four problems can be minimized. This change is inspired by the Fourier transform basis functions. Complex wavelet transform (CWT) [16] is represented in form of complex valued scaling functions and complex valued wavelet functions.

$$\psi_c(t) = \psi_r(t) + j\psi_i(t) \tag{1}$$

$\psi_r(t)$ are real and even and $j\psi_i(t)$ are imaginary and odd. $\psi_r(t)$ and $\psi_i(t)$ form a Hilbert transform pair (90^o out of phase each others) and $\psi_c(t)$ is the analytics signal. Complex scaling function is also defined in similar ways. Projecting the signal onto $2^{j/2}\psi_c(2^j t - n)$, obtain complex wavelet transforms as follows,

$$d_c(j,n) = d_r(j,n) + jd_i(j,n) \tag{2}$$

In the Complex Wavelet domain, analysis depends on two factors, frequency content (which is controlled by scale factor j) and different time (which is controlled by time shift n).

The dual tree CWT employs two real discrete wavelet transforms (DWT); the first DWT gives the real part of the transform while the second DWT gives the imaginary part. The analysis Filter Bank (FB) used to implement the dual-tree CWT is illustrated in Fig 4.

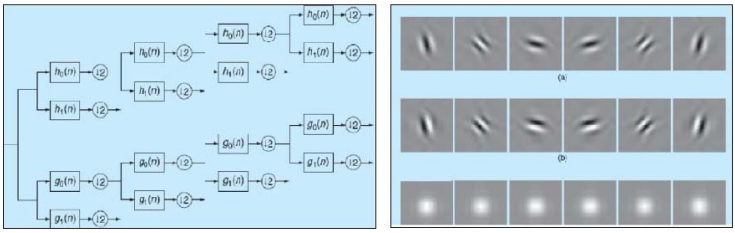


Fig. 4. (left)DTCW filter bank. (right)Typical wavelets associated with the 2-D dual-tree CWT. (a) illustrates the real part of each complex wavelet; (b) illustrates the imaginary part; (c) illustrates the magnitude. Courtesy [16]

To design an overall transform we use two sets of filters. Each set of filter represents real wavelet transform. This overall tranform is approximately analytic. Let $h0(n), h1(n)$ denote the low-pass/high-pass filter pair for the upper FB, and let $g0(n), g1(n)$ denote the low-pass/high-pass filter pair for the lower FB. Denote the two real wavelets affiliated to each of the two real wavelet transforms as $\psi_h(t)$ and $\psi_g(t)$. Filters are designed so that the complex wavelet definition in Equ 1 is approximately estimated. Equivalently, they are designed so that $\psi_g(t)$ is approximately the Hilbert transform of $\psi_h(t)$ [denoted as $\psi_g(t) \approx H\psi_h(t)$]. If the two real DWTs are characterized by the square matrices F_h and F_g, then the dual-tree CWT can be represented by

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_g \\ \mathbf{F}_h \end{bmatrix} \tag{3}$$

If the vector x represents a real signal, then $w_h = F_h x$ represents the real part and $w_g = F_g x$ represents the imaginary part of the dual-tree CWT. The complex coefficients are given by $w_h + jw_g$.

In dual-tree CWT, consider the 2-D wavelet $\psi(x, y) = \psi(x)\psi(y)$ associated with the row-column implementation of the wavelet transform, where $\psi(x)$ is a complex wavelet given by $\psi(x) = \psi_h(x) + j\psi_g(x)$. Obtain $\psi(x, y)$ for the expression,

$$\begin{aligned} \psi(x,y) &= [\psi_h(x) + j\psi_g(x)][\psi_h(y) + j\psi_g(y)] \\ &= \psi_h(x)\psi_h(y) - \psi_g(x)\psi_g(y) \\ &+ j[\psi_g(x)\psi_y(x) + \psi_h(x)\psi_g(y)] \end{aligned} \tag{4}$$

Take the real part of this complex wavelet, then obtain the sum of two divisible wavelets

$$RealPart\psi(x,y) = \psi_h(x)\psi_h(y) - \psi_g(x)\psi_g(y) \tag{5}$$

We perform a 4-level decomposition and compute the wavelet energy (square sum of wavelet coefficients around 5x5 window) of the real-part at each level and use this as feature vector. These parameters were empirically determined to attain highest accuracies for palmprint recognition experiments presented in this paper.

2.2 Local Binary Pattern Features

LBP captures the local level texture variations. Local binary patterns introduced by Ojala et al. [15] use local texture descriptor. In its simplest form, an LBP description of a pixel is created by thresholding the values of the 3x3 neighborhood of the pixel against the central pixel and explicating the result as a binary number. The Local binary pattern (LBP) operator was originally designed as a texture descriptor. The LBP operator attributes a label to every pixel of an image by thresholding the 3x3 neighborhood of each and every pixel value with the center pixel value and assigns binary value (0,1) based on the following equation,

$$I(x,y) = \begin{cases} 0 & \text{if } N(x,y) < I(x,y) \\ 1 & \text{else} N(x,y) \geq I(x,y) \end{cases}$$

where $I(x,y)$ is the center pixel value and $N(x,y)$ is neighborhood pixel value. After thresholding, central pixel value is represented by a binary number (or decimal number) called label. Histogram of these labels is used as texture descriptor. LBP operator is extended to scale invariance and rotation invariance texture operator for images.

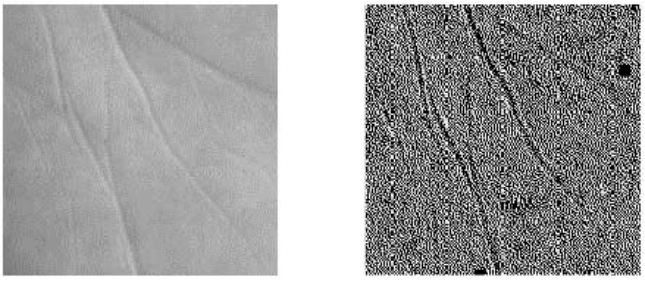


Fig. 5. (left) Extracted 128x128 palm ROI. (right) LBP features on the ROI.

LBP operator deals with textures at different scales using neighborhoods of different sizes. Local neighborhood can be defined using circular neighborhood. Circular neighborhood is a set of evenly spaced sampling points on a circle, whose center is the pixel to be labeled. Radius of circle controls the spatial resolution of operator and number of sampling points controls angular space quantization.

Interpolation is used when a sampling point does not fall in the mediate of a pixel. Notation (P; R) will be used for pixel neighborhoods which contemplates P sampling points on a circle of radius of R.

Fig 5 shows the extracted LBP features of the segmented Region-of-interest on the palm image. To achieve the gray level invariance we subtract the center pixel value with all circular neighborhood pixel values and assume that this difference is independent of center pixel value.

3 Feature Fusion and Matching

Canonical correlation analysis can be defined as the complication of finding two sets of basis vectors, one for $\mathbf{x}$ and one for $\mathbf{y}$, in a way that the correlations between the projections of the variables onto these basis vectors are mutually maximized. Linear combinations $x = \mathbf{x}^T\hat{\mathbf{w}}_x$ and $y = \mathbf{y}^T\hat{\mathbf{w}}_y$ of the two variables is maximized as follows,

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\hat{\mathbf{w}}_\mathbf{x}^\mathbf{T}\mathbf{x}\mathbf{y}^\mathbf{T}\hat{\mathbf{w}}_\mathbf{y}]}{\sqrt{E[\hat{\mathbf{w}}_\mathbf{x}^\mathbf{T}\mathbf{x}\mathbf{x}^\mathbf{T}\hat{\mathbf{w}}_\mathbf{x}]\mathbf{E}[\hat{\mathbf{w}}_\mathbf{y}^\mathbf{T}\mathbf{y}\mathbf{y}^\mathbf{T}\hat{\mathbf{w}}_\mathbf{y}]}} = \frac{\mathbf{w}_x^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}} \tag{6}$$

The maximum of ρ with respect to $\mathbf{w_x}$ and $\mathbf{w_y}$ is the maximum canonical correlation.

$$\begin{cases} E[x_i x_j] = E[\mathbf{w}_{xi}^T\mathbf{x}\mathbf{x}^T\mathbf{w}_{xj}] = \mathbf{w}_{xi}^T\mathbf{C}_{xx}\mathbf{w}_{xj} = 0 \\ E[y_i y_j] = E[\mathbf{w}_{yi}^T\mathbf{y}\mathbf{y}^T\mathbf{w}_{yj}] = \mathbf{w}_{yi}^T\mathbf{C}_{yy}\mathbf{w}_{yj} = 0 \\ E[x_i y_j] = E[\mathbf{w}_{xi}^T\mathbf{x}\mathbf{y}^T\mathbf{w}_{yj}] = \mathbf{w}_{xi}^T\mathbf{C}_{xy}\mathbf{w}_{yj} = 0 \end{cases}$$

The projections onto $\mathbf{w}_x$ and $\mathbf{w}_y$, i.e. x and y, are called canonical variates.

$$a^2 + b^2 = c^2 \tag{7}$$

The covariance matrix between two random variables $\mathbf{x}$ and $\mathbf{y}$ with zero mean is defined as.

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \right] \tag{8}$$

where $\mathbf{C}$ is a block matrix where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-sets covariance matrices of $\mathbf{x}$ and $\mathbf{y}$ respectively and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is the between-sets covariance matrix. The canonical correlations between $\mathbf{x}$ and $\mathbf{y}$ can be found by solving the eigenvalue equations

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho^2\hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho^2\hat{\mathbf{w}}_y \end{cases} \tag{9}$$

where the eigenvalues ρ^2 are the squared canonical correlations and the eigenvectors $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are the normalized canonical correlation basis vectors. The number of non-zero solutions to these equations are limited to the smallest dimensionality of $\mathbf{x}$ and $\mathbf{y}$.

Just one of the eigenvalue equations needs to be solved since the solutions are related by

$$\begin{cases} \mathbf{C}_{xy}\hat{\mathbf{w}}_y = \rho\lambda_x\mathbf{C}_{xx}\hat{\mathbf{w}}_x \\ \mathbf{C}_{yx}\hat{\mathbf{w}}_x = \rho\lambda_y\mathbf{C}_{yy}\hat{\mathbf{w}}_y, \end{cases} \tag{10}$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\hat{\mathbf{w}}_y^T\mathbf{C}_{yy}\hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T\mathbf{C}_{xx}\hat{\mathbf{w}}_x}}. \tag{11}$$

As discussed below, we apply method proposed by [17] to combine the output of DTCW and LBP and maximize the information present in these two feature vectors.

Let the two feature extractors be trained by L training images. Let $A = [a_1, a_2, ..., a_L]$ and $B = [b_1, b_2, ..., b_L]$ be the corresponding outputs of the two extractors, and n1 and n2 be the dimensions of the two outputs, where $n1, n2 \leq L1$.

The covariance matrices for $\mathbf{A}$ and $\mathbf{B}$ are given as $\mathbf{C}_{aa}$ and $\mathbf{C}_{bb}$ respectively. $\mathbf{C}_{ab}$ is the between-set covariance matrix. $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are canonical basis vectors of feature vectors $\mathbf{A}$ and $\mathbf{B}$. a_i and b_i are two feature vectors of image i. Fusion of these two feature vectors is defined as,

$$\mathbf{F}_i = \begin{bmatrix} \hat{\mathbf{w}}_x^T a_i \\ \hat{\mathbf{w}}_y^T b_i \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{w}}_x & 0 \\ 0 & \hat{\mathbf{w}}_y \end{bmatrix}^T \begin{bmatrix} a_i \\ b_i \end{bmatrix} \tag{12}$$

For the matching purpose we use cosine similarity measure. Cosine similarity measure is defined as cosine angle between test image fused feature vector and training images fused feature vector,

$$\arg max_{j\in[1,2\cdots L]} \left(\frac{\mathbf{F}_i^T\mathbf{F}_j}{\|\mathbf{F}_i\|\cdot\|\mathbf{F}_j\|} \right) \tag{13}$$

The maximum value according to Equ(13) is estimated as an authenticated palmprint match.

4 Experiments and Results

We test our algorithm on publicly available CASIA database [12], which contains 5502 palm images of 312 subjects from both left and right palms. There are no pegs to restrict postures and positions of palms hence leading to several palm localization issues. To the best of our knowledge this database is the largest publicly available database in terms of number of subjects [3]. 80% of the images are used as training samples and remaining are used for testing purpose. We resize the images to 128x128.

In first of our experiments, we check for recognition accuracy using DTCW and LBP feature separately. Highest recognition rate of Dual-tree and LBP are 90.8% and 92.6% with 128 and 150 lengths feature vector respectively. Refer Fig 6 which plots the recognition rate vs feature vector length.

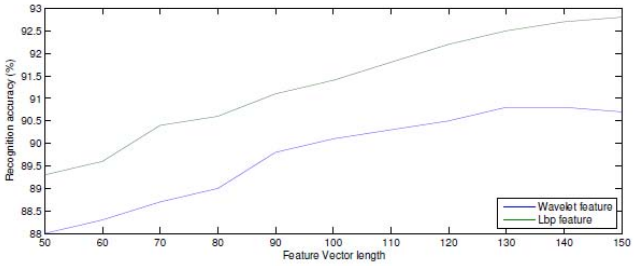


Fig. 6. Recognition performance of individual methods, DTCW and LBP

We perform the second experiment to illustrate the superior performance of CCA based fusion framework over two other widely used fusion methods: Sum rule and Max rule [6]. Fig 7 shows recognition accuracy of three fusion methods used at variable feature vector length. Fusion of these two features improves the recognition rate. Highest recognition rate attained from fusion using Sum rule, Max rule and CCA reached 93.8%, 95.2% and 97.2% respectively with feature vector length 200.

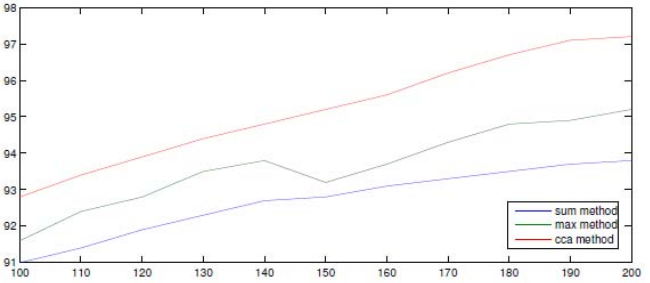


Fig. 7. Accuracy of different fusion methods at variable feature vector length

Hence we infer that local features perform better than global features for palmprint recognition and fusion using Max rule considerably improves the recognition accuracy in comparison with Sum rule. We also establish that Canonical Correlation based fusion outperforms other methods with an EER of 3.2%. See Fig 8.

In [13] DTCW with SVM was used for palmprint recognition on PolyU database which attains an accuracy of 97%. Significant performance of our approach can be noted by comparing the complicacy of CASIA database to PolyU database. As stated in [3 - Table 1 and Table 5], error rates obtained on CASIA are much higher than those obtained on PolyU database. The assumed reasons for this being,

1. The quality of the palm images in CASIA was lesser than those in PolyU as CASIA images were captured using web camera.

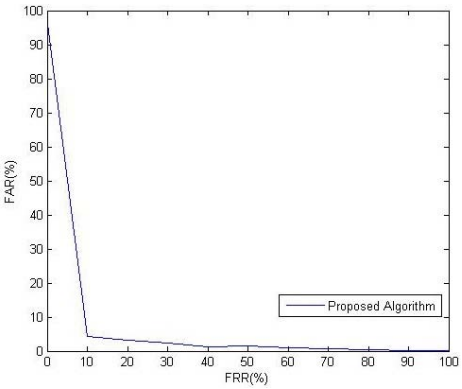


Fig. 8. Receiver Operating Characteristics curve for the proposed method on CASIA palmprint database

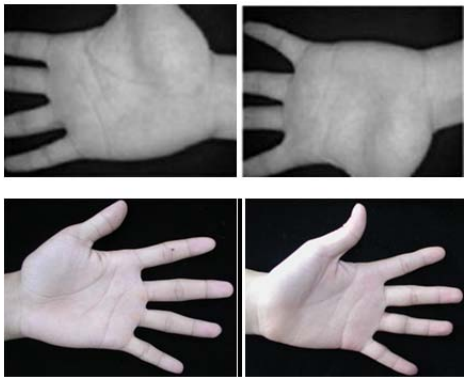


Fig. 9. Palmprint image samples from (top)CASIA palmprint database (bottom)UST hand Image database

2. There are no pegs to restrict postures and positions of palms during data acquisition hence leading to several localization issues like translation, rotation, scaling, varied illumination because of the degrees of freedom.
3. As CASIA is a large database in terms of number of subjects (no. of palms) than PolyU which might degrade the recognition accuracy.

In [14] palmprint identification using Local Binary Pattern and Adaboost was proposed and an Equal Error Rate (EER) of 2% was reported which is much lesser than our EER of 3.2%. Again to justify this, we compare CASIA database to UST hand Image database [20] on which results in [14] were shown. While palmprint images in UST database had a resolution of 24 bits, CASIA palmprints are 8-bit gray-level images. This difference is evident from Fig 8. One more reason might be the selection of discriminative local binary patterns using Adaboost.

In [6], a similar framework based on local and global feature fusion was proposed. Good performance of 97.8% was reported with Sum-rule. We attribute the gain in accuracy on the complicacy of the database used. The internal/private database used in this paper had (a) scanner-based image acquisition system, (b) small in size with 1000 palmprint images from 100 subjects.

5 Conclusion

In this paper, a hybrid feature extraction and fusion framework based on texture information for palmprint recognition is proposed. DTCW captures the global information ensuring scale-invariance and shift-invariance which helps in discriminating between locally similar regions. LBP on the other hand being gray-scale and rotation invariant, captures local fine textures effectively. Improvement in performance is reported by combining these features using three different fusion rules, of which Canonical Correlation based fusion approach is most promising. Through the fusion framework, more than 5% performance gain over single extraction method is reported. In comparison with three other methods [13] [14] [6] which proceed on the similar lines, advantages of our approach are established by supporting reasoning/inferences.

Acknowledgments. Portions of the research in this paper use the CASIA Palmprint Database collected by the Chinese Academy of Sciences Institute of Automation (CASIA). Authors would like to appreciate the efforts in constructing the database.

References

1. Zhang, D., Kong, W.K., You, J., Wong, M.: Online palmprint identification. IEEE Trans. Pattern Anal. Mach. Intell. 25(9), 1041–1050 (2003)
2. Kong, A.W.-K., Zhang, D.: Feature-Level Fusion for Effective Palmprint Authentication. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 761–767. Springer, Heidelberg (2004)
3. Zuo, W., Lin, Z., Guo, Z., Zhang, D.: The multiscale competitive code via sparse representation for palmprint verification. In: Proc. of CVPR, pp. 2265–2272 (2010)
4. You, et al: On hierarchical palmprint coding with multiple features for personal identification in large databases. IEEE Trans. Circuits Syst. Video Tech. (2004)
5. Kong, W., Zhang, D., Li, W.: Palmprint feature extraction using 2-d gabor filters. Pattern Recog. 36(10), 2339–2347 (2003)
6. Pan, et al.: Palmprint recognition using fusion of local and global features. In: Proc. of the Int. Symposium on Intelligent Signal Processing and Communication Systems, pp. 642–645 (2007)
7. Jain, A., Feng, J.: Latent palmprint matching. IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 1032–1047 (2009)
8. Kumar, A., Zhang, D.: Personal authentication using multiple palmprint representation. Pattern Recognition 38, 1695–1704 (2005)
9. Wu, X., Zhang, D., And Wang, K.: Palm line extraction and matching for personal authentication. IEEE Trans. Syst. Man Cybern. Part A 36(5), 978–987 (2006b)

10. Liu, L., Zhang, D., You, J.: Detecting wide lines using isotropic nonlinear filtering. IEEE Trans. Image Process. 16(6), 1584–1595 (2007)
11. Han, C.-C.: A hand-based personal authentication using a coarse-to-fine strategy. Image and Vision Computing 22, 909–918 (2004)
12. CASIA Palmprint database, http://www.idealtest.org/dbDetailForUser.do?id=5
13. Chen, G.Y., Xie, W.F.: Pattern recognition with SVM and dual-tree complex. Image and Vision Computing 25(6), 960–966 (2007)
14. Wang, et al: Palmprint identification using boosting local binary pattern. In: Proceedings of ICPR, pp. 503–506 (2006)
15. Ojala, et al: Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Transactions on PAMI 24(7), 971–987 (2002)
16. Selesnick, et al.: The dual-tree complex wavelet transform. IEEE Signal Processing Magazine 22(2), 123–151 (2005)
17. Sun, Q.-S., et al.: A theorem on the generalized canonical projective vectors. Pattern Recognition 38, 449–452 (2005)
18. Kong, A., Zhang, D., Kamel, M.: Palmprint identification using feature-level fusion. Pattern Recog. 39(3), 478–487 (2006a)
19. Kong, A., Zhang, D.: Competitive coding scheme for palmprint verification. In: Proceedings of the 17th International Conference on Pattern Recognition, pp. 520–523 (2004)
20. UST Hand Image database, http://www4.comp.polyu.edu.hk/~csajaykr/Database/palm/2dhand.html
21. Zhang, et al.: A Comparative Study of Palmprint Recognition Algorithms. ACM Computing Surveys 44(1) (2012)

An Empirical Evaluation of SVM on Meta Features for Authorship Attribution of Online Texts

Hongwei Yao[1], Tieyun Qian[1], Li Chen[2], Manyun Qian[2], and Xueyu Mo[2]

[1] State Key Laboratory of Software Engineering,
Wuhan University, Wuhan, China
[2] Department of Computer Science,
Central China Normal University, Wuhan, China
yaohongwei2012@126.com, qty@whu.edu.cn,
{ccnuchenli,ccnuqianmanyun,guilinmoxueyu}@163.com

Abstract. Authorship attribution (AA) has been studied by many researchers. Recently, with the widespread of online texts, authorship attribution of online texts starts to receive a great deal of attentions. The essence of this problem is to identify a set of features that can capture the writing styles of an author. However, previous studies on feature identification mainly used statistical methods and conducted out experiments on small data sets, i.e., less than 10. This scale is distance from the real application of AA of online texts. In addition, due to the special characteristics of online texts, statistical approaches are rarely used for this problem. As the the performance of authorship identification depends highly on the the combination of the features used and classification methods, the feature sets for traditional authorship attribution needs to be re-examined using machine learning approaches. In this paper, we evaluate the effectiveness of six types of meta features on two public data sets with SVM, a well established machine learning technique. The experimental results show that lexical and syntactic features are the most promising features for AA of online texts. Furthermore, a number of interesting findings regarding the impacts of different types of features on authorship attribution are discovered through our experiments.

Keywords: authorship attribution of online texts, meta features, comparative evaluation.

1 Introduction

Authorship attribution (AA) , also known as authorship classification, has been studied by many researchers [11,27,8,12,6]. It was originally proposed to classify Shakespeare plays, Bronte Sisters' novels, etc. Later on, it was applied to other literary works such as American and English literature and news articles. Recently, AA was applied to various types of online texts such as emails [29,16], blogs [20], forum posts [26] and reviews [17]. We call this theme of studies authorship attribution of online texts. The problem of AA of online texts is useful in many applications, e.g., fake reviewers detection, internet plagiarism and cybercrime investigation.

The number of classes or authors used in the traditional AA is often very small. For example, the corpus used by Escalante et al. [6] comprises documents of 10 authors.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 28–37, 2013.
© Springer International Publishing Switzerland 2013

The number of classes used by Kim et al. [17] is only 8. Recently, Solorio et al. showed that the classification results deteriorated quickly as the number of authors increase from 5 to 100 [26]. However, the number of classes (authors) for AA of online texts can be much larger. For example, in online reviews, a large number of reviewers have written reviews about products and services. The complexity level of AA of online texts is thus becoming extremely high.

The research effects on authorship attribution mainly focus on (1) developing advanced techniques, and (2) extracting effective features. Early studies mainly use statistical univariate methods such as Naïve Bayes (NB) classifier and principle component analysis (PCA). Most modern AA approaches are based on machine learning techniques like decision trees, neural networks, and support vector machine (SVM). Generally, machine learning approaches show better performance than statistical methods. On the other hand, the extraction of writing features has been the essence of AA ever since the earliest work. Almost 1000 features, including length features, richness features, character and lexical features, syntactic features, stylistic features, have been explored in the literature. However, there exist lots of controversies on the effectiveness of different features. The main reason can be due to the various application circumstances and the lack of publicly available data sets. In addition, the performance of authorship analysis depends highly on the the combination of the features used and analytical techniques [32]. It is hard to reach agreements on a best set of features for different approaches. Although Grieve and Halteren conducted quantitative evaluations on a number of features [9,10], these two studies have a number of shortcomings.

- They used the statistical analysis [9] and distance comparison [10] in authorship analysis. Due to the sensitivity to noises, the incapability to deal with the large number of features, and the strong requirements on mathematical assumptions, these techniques are rarely used in modern authorship attribution.
- They carried out experiments on data sets of a relatively small size. The number of authors classified by Grieve [9] and Halteren [10] is 40 and 8, respectively. An analysis on this scale is far from the real application of authorship attribution of online texts.

Given the fact that machine learning techniques achieve high accuracy in AA of online texts [6,26], it is desirable to investigate how the features perform with machine learning approaches for large AA problems. Furthermore, the machine learning technique allows to process a very large number of features. This provides the opportunity to use meta features which enable a closer look at the same type of features on a high level. In this paper, we re-examine the effectiveness of six types of commonly used meta features for AA of online texts. We use two public data sets for online texts. Each of them has a large number of authors, i.e., 62 and 100. The learning technique is support vector machine (SVM). We aim to seek answers to the following questions with empirical evidence.

1. How do the various meta features perform with the widely used machine learning technique like SVM? Are there any meta features consistently better than other features for the large AA problems?
2. Can the attribution task benefit from the combinations of different types of meta features? If this is the case, what is the most promising combination?

2 Related Works

Authorship attribution has received a great deal of attentions in recent years. A variety of approaches have been developed for this problem. Existing methods can be categorized into two main themes. One focuses on finding appropriate features for quantifying the authors' writing style, and the other focuses on developing efficient and effective techniques to perform the classification task.

There is a body of literature examining the effects of different features. The use of function words could date back about half a century ago [22]. Since then, various features have been proposed for modeling writing styles. Existing studies show that the function words [1,3] and rewrite rules [11] might be promising in AA problems. Other features that have been investigated include length features [7,8], richness features [11,19], punctuation frequencies [8], character n-grams [9,12], word n-grams [4,23], POS n-grams [7,13], *k*-ee subtree [17] and topic models [24].

There are also a number of works that study the attribution methods. Mosteller used the Bayesian statistical analysis on function words and obtained good discrimination results [22]. Additional research in recent years focuses exclusively on classification or categorization methodologies, including discriminant analysis [27], PCA [14], exponentiated gradient algorithm [2], neural networks [8,32], multi-layer perceptrons [8], clustering [23], decision trees [28,31], and SVM [5,7,19,12]. In general, machine learning approaches have better performance than statistical methods [32]. In particular, SVM is regarded as one of the best approaches [21,17].

Overall, current surveys on different feature sets are carried out on a small number of authors. There is very limited work on the evaluation of features for the large scale AA problems of online texts. More importantly, previous studies that conduct comparative evaluation on features for AA problem use the statistical analysis [9] and distance comparison [10]. Thorough evaluations using the well established SVM method for large AA problems are still missed in the literature. Therefore, it becomes necessary to make a comprehensive study on how different features affect large authorship attribution of online texts within a SVM framework. That is why we conduct this study.

3 Meta Features

Let $A = \{a_1, a_2, ..., a_k\}$ be a set of *k* authors and $D = \{D_1, D_2, ..., D_k\}$ be *k* sets of documents with D_i being the document set of author $a_i \in A$. Supervised AA builds a model or classifier from the training data and applies it to the test set to determine the author a of each test document d, where a is from A ($a_i \in A$). Each author is treated as a class, and each document is represented as a vector of features. In this paper, we extracted six types of meta features (see below). We do not use any application-specific features such as structural layouts [29] and signature [26] because these features are domain dependant and thus not universally adopted in most of AA problems.

3.1 Length Meta Features

We compute the average document length in terms of word count in one document, the average sentence length in terms of word count in one sentence, and the average

word length in terms of character count in one word, which give us three average length features.

3.2 Character Meta Features

Character n-grams is simple and easily available for any natural language [9]. In this paper, we extract frequencies of n-grams ($n = 1..2$) on the character-level.

3.3 Lexical Meta Features

It is straightforward to view an text article as a bag-of-words, like that has been widely used in topic-based text classification. We represent each article by a vector of word frequencies.

3.4 Syntactic Meta Features

We use four typical content-independent structures including n-grams of POS tags ($n = 1..3$) and rewrite rules [7,13]. The syntactic features are extracted from the parsed syntactic trees. We use the Stanford PCFG parser [18] to generate the grammar structure of sentences in each document.

3.5 Stylistic Meta Features

We derive three stylistic features directly from the raw data. (1) Function words: We use a list of *157* function words in this paper, which is downloaded from $www.flesl.net/Vocabulary/Single-word\_Lists/function\_word\_list.php$. (2) Punctuation frequency [8]: We use *32* common punctuation marks in our experiments. (3) The frequency of each word-length for each article: we get a distribution for k-length word ($1 \leq k \leq 15$).

3.6 Richness Meta Features

Originally, the vocabulary richness functions are used to quantify the diversity of the vocabulary of a text [11,30]. In this paper, we apply the richness metrics to counts of word unigrams, POS tags ($n = 1..3$), and rewrite rules.

4 Experimental Evaluation

All our experiments use the $SVM^{multiclass}$ classifier [15] with default parameter settings. We report classification accuracy as the evaluation metric.

4.1 Experiment Setup

We use two different kinds of online texts. The first one consists of posts from the Chronicle of Higher Education (CHE) [26]. This data set has 100 authors with 16,171 documents. The second one is IMDB data set [25] which contains the IMDB reviews in May 2009. This data set has 62,000 reviews by 62 users (1,000 reviews per user). Both

of the data sets are publicly available upon the request to authors. For CHE data, we conduct experiments on its fixed partition of training (80%) and testing (20%) for all collections. For IMDB, we randomly split training and test documents 5 times, 70% of one author's documents are used for training and the rest 30% for testing. The results are averaged over the 5 splits.

We extract and compute the length, character, lexical, stylistic, and richness meta features directly from the raw data, and we use the Stanford PCFG parser [18] to generate the grammar structure of sentences in each document for extracting syntactic features. We do not remove stop words as some of them are actually function words. We normalize each feature's value to [0, 1] interval by dividing by the maximum value of this feature in the training set. Table 1 shows some statistics on these two data sets.

Table 1. Vocabulary size for different features

Meta Features	Features	Vocabulary Size	
		CHE	IMDB
Length	Avg Doc Len	1	1
	Avg Sent Len	1	1
	Avg Word Len	1	1
Character	Char 1-Gram	1476	2094
	Char 2-Gram	6286	13805
Lexical	Bag of words	34840	195274
Syntactic	POS 1-Gram	63	63
	POS 2-Gram	1575	1917
	POS 3-Gram	12967	21950
	Rewrite Rules	7916	19240
Stylistic	Function Words	157	157
	Punctuation	32	32
	Len k Words	15	15
Richness	Word	6	6
	POS 1-Gram	6	6
	POS 2-Gram	6	6
	POS 3-Gram	6	6
	Rewrite Rules	6	6
	Function Words	6	6

From Table 1, we can see that the two data sets have their own characteristics, and the corresponding classification tasks differ significantly in their difficulty. For example, the number of words in IMDB is greatly larger than that in CHE. In addition, there are more syntactic features in IMDB, indicating the authors intend to use more flexible syntactic structures when writing.

4.2 Results with Single Meta Features

Table 2 presents the results in terms of accuracy with single meta features. It is clear that the lexical and syntactic meta features have the best performance. Especially on IMDB data, their improvements over other meta features are very significant. On the

other hand, both the length and richness meta features are the least successful of all the features. Contradiction to the past research [9] which reported that character n-gram are some of the most accurate techniques in their test, the character n-grams do not have a big enough impact on authorship attribution of online texts. It only ranks the third and the forth among six types of meta data for CHE and IMDB, respectively. Overall, the results for IMDB are much better than those for CHE, which is naturally because CHE contains 100 authors (classes) while IMDB only has 62.

Table 2. Results on single meta features

Meta Features	Acc. (CHE)	Acc. (IMDB)
Length	3.29	1.97
Character	7.32	17.92
Lexical	**17.57**	**47.37**
Syntactic	**12.88**	**50.80**
Stylistic	11.78	12.68
Richness	3.69	2.58

The Performance Ranking of Single Meta Feature

The performance ranking of single meta feature for CHE and IMDB in descending order is {Lexical, Syntactic, Stylistic, Character, Richness, Length}, and {Syntactic, Lexical, Character, Stylistic, Richness, Length}, respectively.

4.3 Results with the Combination of Two Types of Meta Features

Results for Combo_length

It is clear from Table 3 that the combination of length meta feature and any other features performs better than the single length feature. This is intuitive because the length meta feature consists only three dimensions, which are a bit too less for SVM classifier. If we take a closer look, we can also find that the two performance rankings are totally consistent with those for single meta feature. This strongly indicates that the length meta feature has a very slight impact on this task.

Results for Combo_richness

In Table 4, one can see that most of the combinations of richness and other meta features are more successful than richness itself. Almost all of them have a positive change. The only exception is the combination of richness and length on IMDB. It has a negative change. However, the same combination on CHE data set does achieve a significant improvement over its corresponding single richness feature. We also observe that rich_syntac is very close to rich_lex for IMDB but it is higher than rich_lex for CHE. This shows that richness is not a stable combination factor. Its performance varies with the data set and the counterpart meta feature.

Table 3. Results on combo-length meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Len_Char	6.20	+49.94%	17.55	+88.77%
Len_Lex	15.94	+79.36%	40.08	+95.08%
Len_Rich	4.98	+33.54%	2.30	+14.35%
Len_Style	6.45	+48.99%	12.62	+84.39%
Len_Syntac	13.26	+75.19%	50.65	+96.11%

Table 4. Results on combo-richness meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Rich_Char	6.47	+42.97%	15.57	+12.99%
Rich_Len	4.98	+25.90%	2.30	-12.17%
Rich_Lex	12.85	+71.28%	48.37	+94.67%
Rich_Style	7.13	+48.25%	13.42	+80.77%
Rich_Syntac	16.24	+77.28%	48.98	+94.73%

Results for Combo_stylistic

In Table 5, we can see that that the combination of stylistic and other meta features sometimes deteriorates the performance. Another interesting finding is the Style_Lex combo performs significantly better than the Style_Syntac combo on IMDB. However, CHE does not show the same pattern. Indeed, the performance of Style_Lex combo is much worse than lexical meta feature on CHE. Similarly, the Style_Rich combo feature shows the opposite performance change than its single richness feature on CHE and IMDB. These finding infer that the improvement of combination of stylistic with lexical and richness feature is dependent on the characteristic of data sets.

Table 5. Results on combo-stylistic meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Style_Char	6.85	-71.97%	21.89	+42.07%
Style_Len	6.45	-82.64%	12.62	-0.48%
Style_Lex	16.46	+28.43%	54.17	+76.59%
Style_Rich	7.13	-65.22%	13.42	+5.51%
Style_Syntac	21.01	+43.93%	43.98	+71.17%

Results for Combo_character

Table 6 shows the results on combo-character meta features. While other combinations are useful to some extent, the performances of Char_Len and Char_Rich decrease on both the CHE and IMDB data sets. Hence it is not good to combine the character meta feature with either length or richness meta feature.

Table 6. Results on combo-character meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Char_Len	6.20	-18.06%	17.55	-2.10%
Char_Lex	24.45	+70.06%	52.31	+65.74%
Char_Rich	6.47	-13.14%	15.57	-15.09%
Char_Style	6.85	-7.30%	21.89	+18.14%
Char_Syntac	16.49	+55.61%	51.61	+65.28%

Results for Combo_lexical

Table 7 show the results for combo_lexical. The Lex_Syntac combo achieves the most significant improvement on IMDB dataset. This is intuitive since the single lexical and syntactic meta features rank the first and the second by themselves. However, it is a bit surprising that it is Lex_Char rather than Lex_Syntac to be the best for CHE. This hints that some weak meta features may have a very positive impact on AA if it is combined with a strong one.

Table 7. Results on combo-lexical meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Lex_Char	24.45	+28.14%	52.31	+9.44%
Lex_Len	15.94	-10.23%	40.08	-18.19%
Lex_Rich	12.85	-36.73%	48.37	+2.07%
Lex_Style	16.46	-6.74%	54.17	+12.55%
Lex_Syntac	21.80	+19.40%	68.03	+30.37%

Results for Combo_syntactic

Table 8 shows the results for combo_syntactic. All the combo syntactic meta features reach an improvement over the single syntactic meta feature on CHE. However, on IMDB, we note there are some decreases. The reason can be due to that syntactic is already the best single meta feature for IMDB. The combination with a weak meta feature may be harmful to the performance.

The Performance Ranking of Combo Meta Feature

The performance ranking of combo meta features for CHE and IMDB in descending order is {Lex_Char, Lex_Syntac, Syntac_Char, Style_Lex, Syntac_Rich}, and {Lex_Syntac, Style_Lex, Lex_Char, Syntac_Char}, respectively. Note that we only list the combos which perform better than both of the corresponding single meta features.

Table 8. Results on combo-syntactic meta features

	CHE		IMDB	
Meta Features	Acc.	Change	Acc.	Change
Syntac_Char	16.49	+21.89%	51.61	+1.57%
Syntac_Len	13.26	+2.87%	50.65	-0.30%
Syntac_Lex	21.80	+40.92%	68.03	+25.33%
Syntac_Rich	16.24	+20.69%	48.98	-3.72%
Syntac_Style	21.01	+38.70%	43.98	-15.51%

5 Conclusion

In this paper, we adopt a machine learning algorithm, i.e., SVM, to examine the impacts of different meta features on authorship attribution of online texts. We conduct extensive comparative studies in authorship recognition using single and combo meta features on two real world data sets with a large number of classes. We have the following interesting findings. Firstly, the lexical and syntactic meta features are the most promising for AA of online texts, and the the effects of length and richness are trivial. Secondly, the performance of the combination of two types of meta features is dependant on the data and the single meta feature. As some of the combinations deteriorate the performance, one should carefully examine the characteristics of data and the performance of single meta feature before the combination is conducted. Thirdly, our results show that the combination of two strong meta features outperform any of their corresponding individual features, and thus this kind of combination is generally more applicable than that of two weak ones.

Acknowledgements. This research was supported in part by the NSFC project (61272275, 61202036, 61272110, U1135005), and the 111 project(B07037).

References

1. Argamon, S., Levitan, S.: Measuring the usefulness of function words for authorship attribution. In: Literary and Linguistic Computing, pp. 1–3 (2004)
2. Argamon, S., Šarić, M., Stein, S.S.: Style mining of electronic messages for multiple authorship discrimination: First results. In: Proc. of the 9th SIGKDD, pp. 475–480 (2003)
3. Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., Levitan, S.: Stylistic text classification using functional lexical features: Research articles. JASIST 58, 802–822 (2007)
4. Burrows, J.F.: Not unles you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing 7, 91–109 (1992)
5. Diederich, J., Kindermann, J., Leopold, E., Paass, G., Informationstechnik, G.F., Augustin, D.S.: Authorship attribution with support vector machines. Applied Intelligence 19, 109–123 (2000)
6. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proc. of the 49th ACL, pp. 288–298 (2011)
7. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. of the 20th COLING (2004)

8. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering 11, 397–415 (2005)
9. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing 22, 251–270 (2007)
10. van Halteren, H.: Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing 4, 1–17 (2007)
11. van Halteren, H., Tweedie, F., Baayen, H.: Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing 11, 121–132 (1996)
12. Hedegaard, S., Simonsen, J.G.: Lost in translation: authorship attribution using frame semantics. In: Proc. of the 49th ACL, pp. 65–70 (2011)
13. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing 22, 405–417 (2007)
14. Hoover, D.L.: Statistical stylistics and authorship attribution: An empirical investigation. Literary and Linguistic Computing 16, 421–424 (2001)
15. Joachims, T.: Making large-scale support vector machine learning practical. In: Advances in Kernel Methods, pp. 169–184. MIT Press (1999)
16. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/veto meta-classifier for authorship identification - notebook for pan at clef 2011 (2011)
17. Kim, S., Kim, H., Weninger, T., Han, J., Kim, H.D.: Authorship classification: a discriminative syntactic tree mining approach. In: Proc. of the 34th SIGIR, pp. 455–464 (2011)
18. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proc. of the 41st ACL, pp. 423–430 (2003)
19. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proc. of the 21st ICML (2004)
20. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Lang. Resources & Evaluation 45, 83–94 (2011)
21. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Communications of the ACM 49, 76–82 (2006)
22. Mosteller, F.W.: Inference and disputed authorship: The Federalist. Addison-Wesley (1964)
23. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation. In: Proc. of EMNLP, pp. 482–491 (2006)
24. Seroussi, Y., Bohnert, F., Zukerman, I.: Authorship attribution with author-aware topic models. In: Proc. of ACL, pp. 264–269 (2012)
25. Seroussi, Y., Zukerman, I., Bohnert, F.: Collaborative inference of sentiments from texts. In: Proc. of the 18th UMAP, pp. 195–206 (2010)
26. Solorio, T., Pillay, S., Raghavan, S., y Gomez, M.M.: Modality specific meta features for authorship attribution in web forum posts. In: Proc. of the 5th IJCNLP, pp. 156–164 (2011)
27. Stamatatos, E., Kokkinakis, G., Fakotakis, N.: Automatic text categorization in terms of genre and author. Comput. Linguist. 26, 471–495 (2000)
28. Uzuner, Ö., Katz, B.: A comparative study of language models for book and author recognition. In: Proc. of the 2nd IJCNLP, pp. 969–980 (2005)
29. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining email content for author identification forensics. Sigmod Record 30, 55–64 (2001)
30. Yule, G.U.: The statistical study of literary vocabulary. Cambridge University Press (1944)
31. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. In: Proceeding of Information Retrival Technology, pp. 174–189 (2005)
32. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. JASIST 57, 378–393 (2006)

An Empirical Comparison of Discretization Methods for Neural Classifier

M. Gethsiyal Augasta[1] and Thangairulappan Kathirvalavakumar[2]

[1] Department of Computer Applications, Sarah Tucker College, Tirunelveli, 627007, TN, India
augasta@yahoo.com
[2] Department of Computer Science, V.H.N.S.N College, VirudhuNagar, 626001, TN, India
Kathirvalavakumar@yahoo.com

Abstract. Discretization leads the improvement in classification accuracy and generalizes the problem well for further knowledge extraction. As a result researchers have developed various discretization methods for preprocessing the data. This paper provides a survey of existing discretization methods that preprocess the data for datamining. Also the paper evaluates the effectiveness of various discretization methods in terms of better discretization scheme and better accuracy of classification by comparing the performance of some traditional and recent discetization algorithms on six different real datasets namely Iris, Ionosphere, Waveform-5000, Wisconsin breast cancer, Hepatitis Domain and Pima Indian Diabetes. The feedforward neural network with conjugate gradient training algorithm is used to compute the accuracy of classification from the data discretized by those algorithms.

Keywords: Preprocessing, Discretization, Classification, Feedforward Neural networks, Datamining.

1 Introduction

In this modern age, tremendous amount of information have been collected with the help of various technologies like computers, internet, satellites etc. These enormous amounts of data stored in files, databases, and other repositories, could be made as useful for decision making by extracting the interesting hidden information from them. Since people are often unable to extract useful knowledge from such huge datasets, data mining plays a vital role in the knowledge discovery process. Classification is one of the important functions of data mining[1]. Many classification algorithms have been developed for classifying real world datasets. However many classification algorithms such as CLIP[2] and CN2[3] can handle only categorical attribute while others can handle continuous attributes but would perform better on discrete attributes. All algorithms cannot be applied to the real world classification tasks involving continuous attributes. These continuous attributes need to be first discretized. To handle this problem a lot of discretization algorithms have been proposed [4-7].

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 38–49, 2013.
© Springer International Publishing Switzerland 2013

Data discretization is a general purpose pre-processing method that transforms continuous attributes values into finite number of intervals and associates with each interval a numerical discrete value. Alternatively, the discretization reduces the number of distinct values for a given continuous attribute by dividing its range into a finite set of disjoint intervals, and then relates these intervals with meaningful labels[1]. Replacing numerous distinct values of a continuous attribute by a small number of interval labels, leads to a reduced and simplified data representation in data exploration and data mining process. Efficient discretization process of continuous attributes is an essential one because, some learning methods can not handle continuous attributes, the data transformed in a set of intervals are more cognitively relevant for a human interpretation and it makes learning more accurate and faster. Careful selection of an effective discretization method is an important one to produce new and more accurate knowledge.

Classification is one of the data mining problem receiving great attention recently in the database community. The preprocessing steps such as data cleaning, relevance analysis and discretization improve the accuracy, efficiency and scalability of classification process [1]. Classification can be performed using different methods such as decision trees [9], Bayesian classification [10], neural networks [11,12] and genetic algorithms [13]. Among them neural network is used well suited for continuous valued inputs. But the drawbacks of neural network in data mining is that it requires lots of time on thousands of iterations for large training data and it has the poor interpretation ability [1]. One of the most common type of neural network is the feedforward neural network. Conjugate gradient is one of the best neural network learning algorithm. Unlike backpropagation algorithm it does not require the user to specify learning rate and momentum parameters.

The aim of the study is to identify the discretization algorithm which helps to reduce the training time of neural network and also to improve the understandability, accuracy, efficiency and scalability of the classification process. This paper discusses on existing discretization algorithms that preprocess the data for classification in datamining and it compares the performance of some traditional discretization algorithms namely Equal-W, Equal-F, Chimerge and some recent discretization algorithms namely Ex-chi2, CAIM, CACC and DRDS in terms of better discretization scheme, discretization time and better classification accuracy. The feedforward neural network with a scaled conjugate gradient training algorithm [14] is used to compute the classification accuracy of the discretized data. The paper is organized as follows: Section 2 discusses various methods of discretization, Section 3 provides a survey of existing discretization algorithms and discusses its advantages and limitations, Section 4 compares the performance of some existing traditional and recent discretization algorithms based on discretization time, mean number of intervals and classification accuracy by implementing them on six different real datasets namely iris, ionosphere, wave, hepatitis, diabetes and breastw.

2 Discretization Methods

Nowadays, various discretization methods are available for discretizing any continuous data. In general, these discretization methods can be classified into various dimensions such as supervised Vs unsupervised, static Vs dynamic, global Vs local, top-down Vs bottom-up, direct Vs incremental and univariate Vs multivariate. The primary categories of discretization methods are supervised and unsupervised. In the unsupervised methods, continuous ranges are divided into sub-ranges by the user specified parameter, but without the consideration of class information. Equal-W, Equal-F, clustering algorithms like k-means are the examples for unsupervised methods [1]. These methods may not give good results in cases where the distribution of the continuous values is not uniform and where the outliers affect the ranges significantly. Obviously if no class information is available, unsupervised discretization is the only choice. In supervised discretization methods, class information is used to find the proper intervals caused by cut-points. Different methods have been devised to use this class information for finding meaningful intervals in continuous attributes [6,8,15,16]. These supervised discretization methods can be further characterized as error-based, entropy-based or statistics-based according to whether intervals are selected using metrics based on error on the training data, entropy of the intervals, or some statistical measure.

Static methods discretize continuous attributes prior to the learning task. On the contrary dynamic methods discretize continuous attributes when classifier is being built [17]. Global methods are applied once to the entire datasets but the local methods are applied only to the subsets of examples [18]. Bottom-up methods start with complete list of all continuous attributes as cut-points and then remove some of them by merging intervals in each step while the top-down methods start with a empty list of cut-points and add new ones in each step [4,16]. The direct methods require the user to decide the number of intervals, on the other hand incremental methods begin with simple discretization scheme and pass through the refinement process and terminates the discretization at the stopping criterion [5]. Univariate discretization quantifies one continuous attribute at a time while multivariate discretization considers simultaneously multiple attributes [19].

3 Discretization Algorithms

Many discretization algorithms have been developed for the process of data mining in knowledge discovery. Equal-W and Equal-F are the best examples for unsupervised. In Equal-W method the range of values is simply divided into sub ranges of equal extent and in Equal-F method the range is divided into sub ranges containing equal number of examples. The entropy based and Chi-square based methods are the examples for the supervised procedure. The best examples for the supervised top-down algorithms are Information Entropy Maximization [7], CACC [5] and CAIM [4]. These algorithms generally maintain the highest

interdependence between target class and discretized attributes, and attain the best classification accuracy. The famous algorithms in bottom-up methods are chimerge [16], chi2 [20], modified chi2 [21] and extended chi2 [22]. In this section, the literature survey of discretization algorithms based on entropy, chi-square, clustering, class attribute interdependency, range coefficient of dispersion and skewness, clustering etc., are described.

RossQuinlan [22] has developed an algorithm called Iterative Dichotomiser 3 (ID3) to induce best split point in decision trees based on the entropy measure. ID3 employs a greedy search to find potential split-points within the existing range of continuous values. Catlett [23] has proposed a supervised dynamic discretization method that recursively selects cutpoints to maximize Quinlans information gain. It ends when a stopping criterion based on a set of heuristic rules is satisfied. Fayyad and Irani [7] have proposed the entropy based supervised discretization method. The entropy is calculated on the basis of the class label. The best split(s) are found by examining all possible splits and then selecting the optimal split. The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization. Grzymala-Busse [24] divides the continuous attributes in a finite number of intervals with maximum goodness, so that the average-goodness of the final set of intervals is the highest. The algorithm does not need any user-parameter and its complexity is subquadratic.

Kass [25] has proposed the popular top-down discretization algorithm CHAID (chi-squared automatic interaction detection) that uses χ^2 statistic. CHAID starts with one interval for the whole range, based on the p-values which are calculated by comparing the value of the statistic to a χ^2 distribution; it determines the best next split at each step to further split the intervals. Kerber [16] has proposed a bottom-up discretization method chiMerge, based on chi-square (χ^2). It splits or merges the interval based on the consistency of the relative class frequencies of that interval. If the relative class frequencies are not consistent within an interval, it should be split. If two adjacent intervals have the similar relative class frequencies they should be merged. Chi-merge is the most typical bottom-up algorithm. The main drawback of chi-merge is that the user has to provide several parameters such as maximal and minimal intervals. Chi2[20] was proposed based on chi-merge. Chi2 automatically calculates the value of the significance level, but still requires the users to provide an inconsistency rate to stop the merging procedure. Modified chi2[21] replaces the inconsistency checking in chi2 by the quality of approximation after each step of discretization. That makes the modified chi2 as a completely automated method. The extended chi2[22] algorithm determines the predefined misclassification rate from the data itself and also considers the variance in the two adjacent intervals. These modifications make the algorithm to handle the misclassified or uncertain data with better accuracy than the original chi2 algorithm.

k-means [26] is the most popular algorithm in clustering analysis. It is a non-hierarchical partitioning clustering algorithm that is used to discretize continuous valued variables because it calculates continuous distance-based similarity measure to cluster data points. Bradley et al., [27] have proposed a discretization

method in divisive approach which automatically discretize continuous dimension into buckets during OLAP data cube construction and data mining modeling. Ferrandiz and Boull [19] have proposed an extension to the multivariate case, relying on the multivariate definition of discrete neighborhood by means of a non-oriented graph structure. A framework for supervised bipartitioning has been proposed, which applied recursively leads to a new multivariate discretization algorithm. Non-Disjoint Discretization (NDD) forms overlapping intervals for a numeric attribute [28], always locating a value toward the middle of an interval to obtain more reliable probability estimation. It also adjusts the number and size of discretized intervals to the number of training instances, seeking an appropriate trade-off between bias and variance of probability estimation.

Ching et al., [29] have proposed Class-Attribute Dependent Discretizer algorithm (CADD) which uses CAIR criterion [30] to measure the interdependence between classes and the discretized attribute. But it uses a user-specified number of intervals when initializing the discretization intervals and initializes the discretization intervals using a maximum entropy discretization method; such initialization may be the worst starting point in terms of the CAIR criterion. The CAIUR algorithm [31] avoided the disadvantages of the CADD algorithm generating discretization schemes with higher CAIR values, but at the expense of very high-computational cost, making it inapplicable for discretization of continuous attributes that have a large number of unique values. Kurgan and Cios [4] have proposed CAIM algorithm which maximizes mutual class-attribute interdependence and possibly generates the smallest number of intervals for a given continuous attribute. It is superior to the other top-down discretization algorithms since its discretization schemes can generally maintain the highest interdependence between target class and discretized attributes, result to the least number of generated rules, and attain the highest classification accuracy. FCAIM [32] which is an extension of CAIM algorithm, have been proposed to speed up CAIM. The main framework, including the discretization criterion and the stopping criterion, as well as the time complexity between CAIM and FCAIM are all the same. The only difference is the initialization of the boundary point in two algorithms. Tsai et al., [5] have proposed a static, global, incremental, supervised and top-down discretization algorithm called CACC in order to raise the quality of the generated discretization scheme by extending the idea of contingency coefficient and combining it with the greedy method. The main goal and contribution of CACC is to propose a criterion to generate better discretization schemes that can lead to the improvement of accuracy of a learning algorithm.

Cerquides and Lopez [33] have proposed a discretization method based on a distance metric between partitions that can be easily implemented in parallel. This method is very effective and efficient in very large datasets. Adaptive Discretization Intervals (ADI) [34] method uses rules that contain intervals built joining together the low level intervals provided by the discretization algorithm, thus collapsing the search space when it is possible. Also, this representation can use several discretization algorithms at the same time allowing the system

to choose the correct discretization for each problem and attribute. Divina and Marchiori [35] have analyzed experimentally discretization algorithms for handling continuous attributes in evolutionary learning. They consider a learning system that induces a set of rules in a fragment of first-order logic, and introduce a method where a given discretization algorithm is used to generate initial inequalities, which describe subranges of attributes values. Mutation operators exploiting information on the class label of the examples are used during the learning process for refining inequalities. Augasta and Kathirvalavakumar [36] have proposed a new static, global, supervised, incremental and bottom-up discretization algorithm based on coefficient of dispersion and skewness of data range (DRDS). It automates the discretization process by introducing the number of intervals and stopping criterion. The method has two phases. The first phase gets the initial discretization scheme by searching through globally. The second phase, refines the intervals by merging them up to the stopping criterion without affecting the quality of the discretization. The efficiency of the algorithm is shown in terms of better discretization scheme and better accuracy of classification on neural networks.

4 Experimental Evaluation

In this section, the results of some existing discretization algorithms such as Equal-W[1], Equal-F[1], DRDS[36], Chimerge[16], Ex-Chi2[22], CACC[5] and CAIM[4] are evaluted on six well known continuous and mixed mode WEKA's datasets namely iris plants (iris), ionosphere (iono), statlog project heart disease (heart), Pima Indians diabetes (pid), waveform-5000 (wav) and Wisconsin breast cancer (breastw) [37] and compared with each other. The detailed description of the datasets is shown in Table 1. The KEEL software[38] is used to compute

Table 1. Properties of 6 real datasets

	Datasets					
Properties	iris	iono	hea	pid	wav	breastw
No. of classes	3	2	2	2	3	2
No. of examples	150	351	270	768	5000	699
No. of attributes	4	34	13	8	40	9
No. of continuous attributes	4	34	13	8	40	9

the classification accuracy of the discretized datasets. The training and testing examples are selected based on 10-fold cross validation method. In this, the dataset is divided into ten disjoint groups of equal size. The training procedure for each data set is repeated 10 times, each time with nine partitions as training data and one partition as test data. All the reported results are the average of the outcome of the 10 separate tests.

4.1 Comparison of Discretization Schemes

The comparison results of seven discretization schemes on six datasets are shown in Table 2. The discretization schemes Equal-W and Equal-F are two unsuper-

Table 2. Comparison of the seven discretization schemes on six datasets

Criterion	Discretization Methods	Datasets					
		iris	iono	heart	pid	wav	breastw
Mean Number of Intervals	Equal-W	4.0	20.0	10.0	14.0	20.0	14.0
	Equal-F	4.0	20.0	10.0	14.0	20.0	14.0
	DRDS	5.75	5.1	5.0	10.8	12.4	4.0
	Chimerge	3.5	21.4	7.8	25.6	28.5	4.6
	Ex-chi2	7.5	8.8	2.3	20.0	12.2	3.3
	CACC	3.0	4.3	6.4	11.2	18.1	2.0
	CAIM	3.0	2.0	2.0	2.0	3.0	2.0
Discretization Time (s)	Equal-w	0.02	1.72	0.12	0.33	9.06	0.26
	Equal-F	0.03	1.84	0.12	0.33	9.33	0.27
	DRDS	0.09	0.64	0.31	1.74	35.7	0.15
	Chimerge	0.09	4.28	0.39	0.94	64.33	0.66
	Ex-chi2	0.11	11.11	1.68	3.23	136.0	1.91
	CACC	0.08	3.62	0.22	0.90	61.41	0.58
	CAIM	0.08	3.43	0.20	0.80	52.38	0.58

vised top down methods, Chimerge, Extended Chi2 and DRDS are three bottom-up approach methods and CACC and CAIM are two new supervised top-down methods. The unsupervised algorithms such as Equal-W, Equal-F and the supervised algorithm chimerge require the user to specify the number of discrete intervals. Other supervised algorithms apply their own criteria to generate an appropriate number of discrete intervals. The CAIM method applies a criteria that can maintain the highest interdependence between target class and discretized attributes, result to the least number of generated rules. CACC uses the criteria based on the idea of contingency coefficient to generate better discretization schemes that lead to the improvement of accuracy of a learning algorithm. DRDS applies the criteria based on the dispersion and skewness of the data that can generates better discretization scheme in minimum discretization time with better classification accuracy.

The main goal of one discretization algorithm should be, generating the better discretization scheme automatically in minimum discretization time that can improve the accuracy and efficiency of learning algorithm. Table 2 shows the number of discrete intervals obtained in this study. According to the order of minimum mean number of intervals generated by the discretization algorithms considered in this study, the algorithms can be sorted as CAIM, DRDS, CACC, Ex-Chi2, Equal-w, Equal-F and Chimerge. It shows that the top-down supervised algorithm CAIM, followed by the bottom-up supervised algorithm DRDS performs outstandingly in the generation of discrete intervals than the other top-down algorithms and other bottom-up algorithms considered in this study.

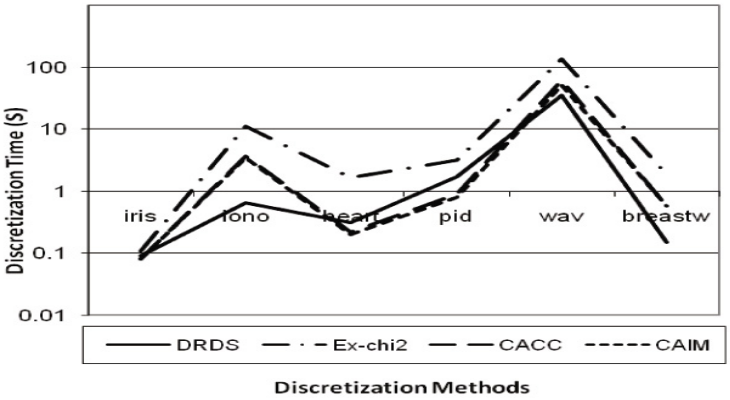


Fig. 1. Comparison of the discretization time of CAIM, CACC, DRDS and Ex-Chi2 on six datasets

Regarding the discretization time, the unsupervised methods are the fastest since they are not considered any class related information. Fig. 1 compares the discretization time of the algorithms which require no parameters from user. Here the discretization time of DRDS is smaller than the other bottom-up method Ex-chi2 for all datasets and smaller than the other top-down methods for three datasets. While comparing the mean discretization time of each algorithm considered in this study for the six experimental datasets, the bottom-up method DRDS uses minimum discretization time than all other algorithms in the study as it need not requires sorting. Normally the bottom-up methods require more execution time to check the merged inconsistency in every step[5], but the bottom-up method DRDS requires less discretization time due to its low computational cost [44].

The classification performance of neural networks on the discretized data of each algorithm considered in our study are evaluated using KEEL software. The accuracy is computed for the six datasets and the results are tabulated in

Table 3. The accuracy obtained by MLP-CG on six datasets

Discretization Type	Discretization Methods	Datasets					
		iris	iono	heart	pid	wav	breastw
Unsupervised	Equal-w	96.6	89.7	77.4	74.1	74.3	94.1
	Equal-F	95.3	84.6	73.7	71.9	79.1	95.7
Supervised	Chimerge	**96.0**	89.4	57.8	65.1	78.3	**96.3**
	Ex-chi2	93.3	64.1	55.5	72.6	77.4	95.1
	DRDS	**96.0**	90.1	**80.7**	**74.9**	**81.3**	95.4
	CACC	93.0	**90.3**	79.3	72.9	80.2	95.1
	CAIM	94.6	89.5	77.0	72.1	78.1	94.9
No discretization	Continuous Data	96.6	87.4	73.7	73.9	86.4	93.9

Table 3. Here the DRDS achieves the highest classification accuracy for four datasets out of 6 among all other supervised discretization methods. Comparing the bottom-up methods namely chimerge, Ex-chi2 and DRDS, the bottom-up method DRDS achieves an equal or high accuracy for all datasets except for the breastw dataset and comparing DRDS with top-down methods CACC and CAIM, DRDS achieves a high or closer accuracy for all datasets. Table 3 also compares classification accuracy of MLP-CG on discretized data of all algorithms in the study with the classification accuracy of MLP-CG on continuous data. Generally the neural network performs well on continuous data. The comparison results in this study show that the neural networks also can be trained with discretized data to achieve the better classification accuracy. In particular, the discretized data of DRDS achieves better classification accuracy than continuous data for the maximum experimental datasets.

People often refused to choose the neural network for classifying large datasets, since it requires long training time. Normally the data discretized with unsupervised discretization algorithms or with some supervised algorithms requires long training time [1]. But the comparison results in Table 4 shows that data discretized by the DRDS achieves the highest classification accuracy with minimum learning time. Table 4 compares the classification performances of neural

Table 4. The accuracy obtained by BPN on six datasets

Methods	Discretiztion Methods												No Discretization			
	DRDS				Equal-w				CAIM				Continuous Data			
Datasets	Epochs	time	acc	mse	Epochs	time	acc	mse	Epochs	time	acc	mse	Epochs	time	acc	mse
iris	100	0.18	96	0.007	100	0.18	92	0.03	100	0.15	94.7	0.009	119	0.19	96.7	0.009
iono	21	0.53	93.7	0.001	249	2.9	91.4	0.001	167	2.23	93.7	0.001	251	1.69	93.1	0.001
heart	165	0.59	83	0.009	480	1.43	76.3	0.009	570	1.47	81.4	0.009	200	0.48	78.6	0.07
pid	100	0.54	75.6	0.03	100	0.56	73.6	0.07	100	0.41	76.5	0.05	100	0.25	66.7	0.15
wav	100	34.5	80.5	0.007	100	24.7	78.3	0.07	100	13.94	77.1	0.03	100	6.8	81.4	0.07
breastw	39	0.34	95.6	0.001	37	0.33	94.3	0.001	552	3.78	91.9	0.001	200	0.91	96.6	0.009

network on continuous and discretized data of outstanding discretization algorithms of this study namely DRDS (bottom-up supervised), CAIM (top-down supervised) and Equal-W(unsupervised) in terms of number of training epochs, training time in seconds, testing accuracy (acc) and mean squared error (mse). The classification performances of the discretized data of these algorithms are evaluated using the Backpropagation neural networks (BPN) [1] by implementing the experiment in JDK1.5 . The three layer feedforward neural network is trained with the discretized training patterns of the dataset using the backpropagation algorithm. This algorithm uses momentum (μ) as 0.5 for all datasets and the learning rate (λ) as 0.1 for four datasets namely iris, cancer, heart and diabetes and 0.9 for two datasets namely ionosphere and wave. Number of input neurons equals the number of attributes in the dataset, Number of output neurons equals the number of target classes and the number of hidden neurons are selected depends on the complexity of the problem. The network is trained

until the error converges to predetermined mean squared error or the prespecified maximum number of iterations has expired, whichever is earlier. Results show that the discretized data of DRDS requires minimum number of epochs for training, minimum training time and achieves higher classification accuracy than CAIM for all datasets and obtains higher or similar performance as in the results of continuous data. Neural networks with discretized data not only improves the classification accuracy but also leads the improvement in the explanation ability of neural networks[1] as the discretization always generalizes the problem well and paves the way to extract knowledge easily in the form of simple rules.

As a summary the discretization algorithm, which provides better discretization scheme in minimum discretization time and achieves better classification accuracy on neural network, can be selected as a preprocessor for neural classifier as it helps to improve the explanation ability of neural networks by the process of rule extraction.

5 Conclusion

In this paper a survey of discretization algorithms which preprocess the data for classification in datamining have been specified. In addition some of the existing unsupervised top-down, supervised top-down and supervised bottom-up discretization algorithms methods are compared on six real datasets with regard to their discretization performance and classification performance on neural networks. The unsupervised methods such as Equal-W and Equal-F achieve faster discretization but fails to improve the classification accuracy as they do not consider the class information for generating the discrete intervals. The comparison results in Table 3 show that among the supervised algorithms, the discretization algorithms CAIM and DRDS of top-down and bottom-up respectively achieve better discretization scheme in minimum discretization time. Though the CAIM method obtains smaller mean number of intervals than DRDS, comparison results in Table 4 show that the DRDS method achieves the discretization goal by obtaining the better classification accuracy in minimum training time using neural networks. Also the discretized data of DRDS performs outstandingly on neural networks than the classification performance of neural networks on continuous data.

Generally neural network is considered as black box as it is very difficult to understand how an ANN has solved a problem. This limitation can be avoided by using the discretization algorithm such as DRDS as a preprocessor for neural classifier as discretization always generalizes the problem well for further knowledge extraction. In a nutshell, the comparative study on discretization algorithms for classification problems using neural networks indicate that the algorithms which consider class information as well as the dispersion and skewness of data for generating discrete intervals as in DRDS are the currently available best discretization algorithms of feedforward neural network for classifying large datasets.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2001)
2. Cios, K.J., Kurgan, L.A.: CLIP 4: Hybrid inductive machine learning algorithm that generates inequality rules. Information Science 163, 37–83 (2004)
3. Clark, P., Niblett, T.: The CN2 algorithm. Machine Learning 3, 261–283 (1989)
4. Kurgan, L.A., Cios, K.J.: CAIM Discretization Algorithm. IEEE Trans. on Knowledge and Data Engineering 16, 145–152 (2004)
5. Tsai, C.J., Lee, C.I., Yang, W.P.: A Discretization algorithm based on Class-Attribute Contingency Coefficient. Information Sciences 178, 714–731 (2008)
6. Butterworth, R., Simovici, D.A., Santos, G.S., Machado, L.O.: A Greedy Algorithm for supervised discretization. Biomedical Informatics 37, 285–292 (2004)
7. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proc. of Thirteenth Int. Conf. on Artificial Intelligence, pp. 1022–1027 (1993)
8. Liu, H., Hussain, F., Tan, C.L., Dash, M.: Discretization: An Enabling Technique. Data Mining and Knowledge Discovery 6, 393–423 (2002)
9. Cohen, S., Rokach, L., Maimon, O.: Decision-tree instance-space decomposition with grouped gain-ratio. Information Sciences 177, 3592–3612 (2007)
10. Yager, R.R.: An extension of the naive Bayesian Classifier. Information Sciences 176, 577–588 (2006)
11. Kaikhah, K., Doddmeti, S.: Discovering trends in large datasets using neural network. Applied Intelligence 29, 51–60 (2006)
12. Ozekes, S., Osman, O.: Classification and prediction in data mining with neural networks. Electrical and Electronics Engineering 3, 707–712 (2003)
13. Dam, H., Abbass, H.A., Lokan, C., Yao, X.: Neural based learning classifier systems. IEEE Trans. on Knowledge and Data Engineering 20, 26–39 (2008)
14. Moller, F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6, 525–533 (1990)
15. Kurgan, L.A., Cios, K.J.: Fast Class-Attribute Interdependence Maximization (FCAIM) Discretization Algorithm. In: Proc. of Int. Conf. on Machine Learning and Applications, pp. 30–36 (2003)
16. Kerber, R.: ChiMerge: Discretization of numeric attributes. In: Proc. of Ninth Int. Conf. on Artificial Intelligence, pp. 123–128 (1992)
17. Wu, Q., Bell, D.A., McGinnity, M., Prasad, G., Qi, G., Huang, X.: Improvement of Decision Accuracy Using Discretization of Continuous Attributes. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 674–683. Springer, Heidelberg (2006)
18. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)
19. Ferrandiz, S., Boullé, M.: Multivariate discretization by recursive supervised bipartition of graph. In: Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS (LNAI), vol. 3587, pp. 253–264. Springer, Heidelberg (2005)
20. Liu, H., Setiono, R.: Feature selection via discretization. IEEE Trans. on Knowledge and Data Engineering 9, 642–645 (1997)
21. Tay, F., Shen, L.: A modified chi2 algorithm for discretization. IEEE Trans. on Knowledge and Data Engineering 14, 666–670 (2002)
22. Su, C.T., Hsu, J.H.: An extended chi2 algorithm for discretization of real value attributes. IEEE Trans. on Knowledge and Data Engineering 17, 437–441 (2005)

23. Catlett, J.: On changing continuous attributes into ordered discrete attributes. In: Kodratoff, Y. (ed.) EWSL 1991. LNCS, vol. 482, pp. 164–177. Springer, Heidelberg (1991)
24. Grzymała-Busse, J.W.: Three strategies to rule induction from data with numerical attributes. In: Peters, J.F., Skowron, A., Dubois, D., Grzymała-Busse, J.W., Inuiguchi, M., Polkowski, L. (eds.) Transactions on Rough Sets II. LNCS, vol. 3135, pp. 54–62. Springer, Heidelberg (2004)
25. Kass, G.: An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29, 119–127 (1980)
26. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
27. Bradley, Fayyad, Reina: Scaling EM (Expectation-Maximization) Clustering to Large Databases. Technical Report MSR-TR-98-35, Microsoft Research (1998)
28. Yang, Y., Webb, G.I.: Proportional k-Interval Discretization for Naive-Bayes Classifiers. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 564–575. Springer, Heidelberg (2001)
29. Ching, J.Y., Wong, A.K.C., Chan, K.C.C.: Class-dependent discretization for inductive learning from continuous and mixed mode data. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 641–651 (1995)
30. Cios, K.J., Pedrycz, W., Swiniarski, R.: Data Mining Methods for Knowledge Discovery. Kluwer (1998), http://www.wkap.nl/~book.htm/0-7923-8252-8
31. Huang, W.: Discretization of Continuous Attributes for Inductive Machine Learning, masters thesis, Dept. Computer Science, Univ. of Toledo, Ohio (1996)
32. Kurgan, L.A., Cios, K.J.: Fast Class-Attribute Interdependence Maximization (FCAIM) Discretization Algorithm. In: Proc. of Int. Conf. on Machine Learning and Applications, pp. 30–36 (2003)
33. Cerquides, J., Lopez de Mantaras, R.: Maximum a posteriori tree augmented naive bayes classifiers. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 73–88. Springer, Heidelberg (2004)
34. Bacardit, J., Garrell, J.M.: Evolving Multiple Discretizations with Adaptive Intervals for a Pittsburgh Rule-Based Learning Classifier System. In: Cantú-Paz, E., Foster, J.A., Deb, K., Davis, L., Roy, R., O'Reilly, U.-M., Beyer, H.-G., Kendall, G., Wilson, S.W., Harman, M., Wegener, J., Dasgupta, D., Potter, M.A., Schultz, A., Dowsland, K.A., Jonoska, N., Miller, J., Standish, R.K. (eds.) GECCO 2003. LNCS, vol. 2724, pp. 1818–1831. Springer, Heidelberg (2003)
35. Divina, F., Marchiori, E.: Handling continuous attributes in an evolutionary inductive learner. IEEE Trans. on Evolutionary Computation 9, 31–43 (2005)
36. Gethsiyal Augasta, M., Kathirvalavakumar, T.: A new Discretization algorithm based on Range coefficient of Dispersion and Skewness for neural networks classifier. Applied Soft Computing 12, 619–625 (2012)
37. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F.: KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13(3), 307–318 (2009)
38. http://weka.wikispaces.com/Datasets

Using a Normalized Score Multi-Label KNN to Classify Multi-label Herbal Formulae

Verayuth Lertnattee, Sinthop Chomya, and Chanisara Lueviphan

Faculty of Pharmacy, Silpakorn University, Sanamchandra Palace, Muang, Nakorn Pathom 73000, Thailand
verayuths@hotmail.com, {verayuth,sinthop,chanisara}@su.ac.th

Abstract. The popularity of herbal medicines has greatly increased in worldwide countries over recent years. Herbal formula is a form of traditional medicine where herbs are combined to heal patient to heal faster and more efficiency. Herbal formulae can be divided into categories. Some formulae can be classified as more than one category. The categories are usually based on indications of herbs in formulae. To support experts for classifying a formula to one or more therapeutic categories, the normalized score multi-label k-nearest neighbors (NSML k-NN) algorithm, is proposed for multi-label herbal formulae classification. The k-NN classifiers with several term weight schemes are explored. The normalized scores are calculated. The values of k, strategies to assign categories are investigated to adjust the decision for multi-label herbal formulae. The experiment is done using a mixed data set of herbal formulae collected from the Natural List of Essential Medicine and the list of common household remedies for traditional medicine. Moreover, a set of well-known commercial products are used for evaluating the effectiveness of the proposed method. From the results, the NSML k-NN is an efficient method to classify multi-label herbal formulae.

Keywords: Multi-label document, text classification, text categorization, herbal formula, k-NN classifier.

1 Introduction

Origins of many traditional treatments in Thailand can be traced to India. The derivation has been diversified throughout many cultures since then [1]. Herbs are natural products that have been used safely for thousands of years to promote healing in patients. The popularity of herbal medicines has greatly increased in worldwide countries over recent years since the World Health Organization (WHO) suggested its member countries to use folk healing practices and herbal medicines as part of the basic public health projects [2]. They should be taken with caution, and careful consideration of the dosage recommended. Traditional herbal formulae can be usually characterized by the use of several herbs. Various patterns of combinations from these herbs, can be applied on a disease. According to Thai traditional medicine, herbal formulae can be divided into categories.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 50–61, 2013.
© Springer International Publishing Switzerland 2013

Some formulae can be classified as more than one category. The categories are usually based on indications of herbs in formulae. A combination of herbs may cause a formula has several categories. These categories may be arranged in flat and/or hierarchy. When the categories are arranged in flat, several main indications of the herbal formula can be applied to patients. In a complex situation, a formula is classified with one main category (or more) and a set of subcategories under the main category. With observation from human, it is hard to discover the combinational patterns of herbs in formulae. Nowadays, several data mining techniques, i.e., classification, clustering, association rules and, etc., have been developed and applied on several types of data. However, only few research works have applied these techniques on herbal information. In this paper, we apply text classification concept to categorize herbal formulae into therapeutic categories. However, many formulae have more than one category, even in a single component formula. Therefore, the normalized score multi-label k nearest neighbors classifier is introduced to the multi-label herbal formulae. To the best of our knowledge, there is no research work contributes to classify multi-label herbal formulae. Performance of the proposed method is investigated on a set of herbal formulae found in the National List of Essential Drugs on both combine and single herbal formulae. Furthermore, a set of commercial traditional herbal formulae products is used for evaluating our proposed method. In the rest of this paper, section 2 presents herbal formulae. The concept of text categorization for herbal formulae is given in section 3. Section 4 introduces the multi-label herbal formula classification by the proposed method. The experimental settings are described in section 5. In section 6, a number of experimental results are given. A conclusion and future work is made in section 8.

2 Herbal Formulae

As opposed to the western medicine, herbs are often used in formulas instead of being used singularly in larger amounts. Formulas allow us to blend herbs to enhance their positive effects (indications) and reduce or eliminate any negative effects (e.g., side effect and/or adverse effects) they may have. These formulas take a very long period of practice to master. Furthermore, the one benefit herbal medicine is that it allows practitioners to set formulas to match each patient and their signs and symptoms exactly. Instead of having a standard formula for a particular condition you can increase the clinical effectiveness of the herbs through this tailoring. For the patient this ideally means faster results with fewer side effects. In Thailand, some items of herbal medicines were placed on the National List of Essential Medicines (EM) in 1999, as part of an effort to promote the use of herbal medicines. Moreover, a list of common household traditional medicines was set. Two types of herbal medicine are placed. The first one is a set of composite well-known herbal formulae. The other one is a set of single herbs that can be used as single herbal formulae or can be combined to composite herbal formulae. According to the National List of Essential Medicine 2013 (the current version), two types of herbal medicines are placed, i.e., 1)

Composite herbal medicines which are composed of several herbs, have been used traditionally and widely by the people for a long time. 2) Herbal medicines which have been developed from a single herb with evidence indicating its safety for use in humans. In the current version, eight therapeutic categories are placed, i.e., cardiovascular, gastrointestinal, gynecologic, antipyretic, respiratory, blood tonic, musculoskeletal and elementary balance. An herbal formula (single or composite), is belong to one or more categories.

3 Multi-label Text Classification and Text Classifiers

With the increasing availability of online information, text classification (TC) turns into the important techniques by using machine learning. The objective of machine learning is to learn classifiers from examples which perform the category assignments automatically. This type of learning is induction-based supervised concept learning or just supervised learning. The supervised learning is the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data contained within a database [3]. Therefore, text classification falls within the machine learning paradigm and data mining. The definition of TC is the activity of labeling natural language texts with thematic categories from a predefined set [4]. Several researches on TC contributed to single-label classification. However, this paper focuses on multi-label herbal formulae classification. Classification techniques have been developed in a variety of learning techniques such as probabilistic models [5], example-based models (e.g., k nearest neighbors, k-NN) [6], linear models [7], support vector machine [8] and so on. The k-NN algorithm is one on the most popular method for TC. For the rest of this section, details of the k-NN text classifier for herbal formulae and multi-label herbal formulae classification are given.

3.1 The *k*-Nearest Neighbor Text Classifier for Herbal Formulae

In the k nearest neighbors text categorization, a document is represented by a vector using a vector space model with a bag of words (BOW) [9]. A set of words in the task of herbal formulae classification is the set of components to combine into a formula of traditional herbal medicine. These components may be terms of crude drugs or natural sources for the crude drugs, i.e., plants, animals and elements. The simplest and popular one is applied term frequency (tf) and inverse document frequency (idf) in the form of $tf \times idf$ for representing a document. In this work, a herbal formula is used instead of a document. The tf means a weight of a crude drug in gram or milliliter in a formula. In a vector space model, given a set of herbal formulae $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|}\}$, a formula f_j is represented by a formula vector $\boldsymbol{f}_j = \{w_{1j}, w_{2j}, ..., w_{|\mathcal{T}|j}\} = \{tf_{1j} \times idf_1, tf_{2j} \times idf_2, ..., tf_{|\mathcal{T}|j} \times idf_{|\mathcal{T}|}\}$, where w_{ij} is a weight assigned to a term (crude drug/natural product sources) t_i in a set of terms ($\mathcal{T}$) of the herbal formula. In this definition, tf_{ij} is term frequency of a term t_i in a formula f_j and idf_i is inverse document (formula) frequency, defined as $log(|\mathcal{F}|/df_i)$. The idf can be applied to eliminate the impact

of frequent terms that exist in almost all documents (formulae). Here, $|\mathcal{F}|$ is the total number of formulae in a collection and df_i is the number of formulae which contain the term t_i. Besides term weighting, normalization is another important factor to represent a document or a class. Term weighting described above can also be applied to both a training document and a test (query) document. Once a training vector and a query vector have been constructed, the similarity between these two vectors can be calculated. The most popular one is cosine similarity [10]. To assign a category (categories) to a new formula, a system finds the k nearest neighbors among the training formulae, and uses the categories of the k nearest neighbors to calculate a category (categories) for the new formula. One of the advantages of the k-NN algorithm is that it can handle non-linear problems. However, its performance depends on tuning the three parameters, i.e., term weighting, appropriate value of the parameter k and strategies of applying similarity function.

3.2 Multi-label Herbal Formulae Classification

For a single-label TC, a document is assigned only one category. Two approaches of classification are utilized to handle single-label classification, i.e., binary classification or multi-class classification. However, the problems in real work usually fall into the problem of multi-label text categorization, where each text document is assigned to one or more categories. Existing methods for multi-label classification can be divided into two main methods, i.e., problem transformation methods and algorithm adaptation methods [11]. Problem transformation methods can be defined as methods that transform the multi-label classification problem either into one or more single-label classification problems or regression problems. This is the same solution to solve the problem of single-label classification. However, the binary classification is based on the assumption of label independence. Therefore, during its transformation process, this method ignores label correlations that exist in the training data. Due to this information loss, predicted label sets from the binary classification are likely to contain either too few or too many labels, or labels that would never co-occur in practice [12]. With some limitations of binary classification, some extensions had been done such as [12,13] to provide better performance of classification. For algorithm adaptation methods, they can be defined as methods that extend or modify specific learning algorithms in order to handle multi-label data directly. The examples for these methods are shown in [14]. Some multi-label k-NN classifiers were applied on TC [15,16]. To the best of our knowledge, there is no research work applied the concept of TC to classify multi-label herbal formulae.

4 Classification of Multi-label Herbal Formulae by the Normalized Score Multi-label *k*-NN

As introduce in the section 3.1, the performance of the k-NN depends on the three parameters, i.e., term weighting, appropriate value of the parameter k and strategies of assigning categories. The tuning of these parameters are described.

4.1 Term Weighting

Performance of a classifier strongly related to term weighting. In k-NN, three components of term weighting are usually considered, i.e., term frequency, collection frequency and normalization components [9]. For term frequency component, three popular term frequency patterns are binary frequency (1 for terms present and 0 for term absent), normal (raw) term frequency, which is mentioned in section 3.1, and the normalized term frequency. An augmented term frequency which is defined as $0.5 + 0.5\frac{tf}{max\ tf}$, is one of popular normalized term frequency. The value of the augmented term frequency is lies between 0.5 and 1.0. For collection frequency component, the *idf* which is described in section 3.1 is usually used. The normalization component is usually applied when the size of formulae is diverse. For herbal formulae, the size is varied from one component to more than 20 components. For this case, a normalization component is important. It is usually applied cosine normalization where each term weight is divided by a factor representing Euclidian vector length.

4.2 The k Nearest Neighbors

The k-NN classifier ranks the test formula's neighbors among the training vectors and uses the category labels of the k most similar neighbors to predict categories of the test formula. In traditional k-NN, the value k is fixed and usually determined experimentally. If the k is too large, big classes (a lot of members in classes) may dominate small ones. Incorrect categories may be assigned for multi-label classification. In the opposite, if k is too small, the advantage of this algorithm to make use of many experts will not be presented. Moreover, in multi-label classification, the test formula may not be assigned all categories it should be.

4.3 Strategies for Assigning Categories

In k-NN algorithm, the most popular on similarity, i.e., cosine similarity [9,10], which can be calculated by the dot product between these two vectors. In case of both vectors are normalized into the unit length, the value of similarity of between the two vectors is in range of 0 and 1. When the k nearest neighbors are set, several strategies could be taken to predict the category of a test formula. Two strategies are widely used are listed as follow.

$$C(f_l) = \arg\max_{c_k \in C} \sum_{f_j \in kNN} z(f_j, c_k) \tag{1}$$

$$= \arg\max_{c_k \in C} \sum_{f_j \in kNN} sim(\boldsymbol{f_l}, \boldsymbol{f_j}) z(f_j, c_k) \tag{2}$$

$$= \arg\max_{c_k \in C} \arg\max_{f_j \in kNN} sim(\boldsymbol{f_l}, \boldsymbol{f_j}) z(f_j, c_k) \tag{3}$$

$$= \arg\max_{c_k \in C} \frac{\sum_{f_j \in kNN} sim(\boldsymbol{f_l}, \boldsymbol{f_j}) z(f_j, c_k)}{\sum_{f_j \in kNN} z(f_j, c_k)} \tag{4}$$

where f_l is a test formula, f_j is one of the neighbors (kNN) in the training set, $z(f_j, c_k) \in \{0, 1\}$ indicates whether f_j belongs to class c_k in the set of classes C, and $sim(\boldsymbol{f_l}, \boldsymbol{f_j})$ is the similarity function between $\boldsymbol{f_l}$ and $\boldsymbol{f_j}$. For single-label classification, the equation 1 means the prediction will be the category that has the largest number of members in the k nearest neighbors. The equation 2 expresses the category which has maximal sum of similarity (score), will be assigned. This strategy is thought to be useful than the equation 1 and is more widely used.

In this paper, a multi-label formula is taken into account. A ranking categorization is applied for this work. Given a formula f_l, a system ranks the categories in according to their estimated similarity to formula f_l . A ranked list of possible therapeutic categories will be considered. The problem of multi-label multiclass classification is how many categories belong to this formula. In this paper, these three parameters are investigated, i.e., term weighting, k nearest neighbors and strategy to assigning categories. For term weighting, the normal term frequency, augmented term frequency and inverse document frequency are used as components along with cosine normalization. In order to set the value of k, heuristics observed from the training set is used, i.e., therapeutic categories of a herbal formula are usually less than or equal to three. The value of k is calculated from mn, where m is the expected maximum categories for a formula and n is the expected number of experts per category. If both m and n are set to 3, then the value of k is 9. Therefore, the value of 10 may be used. For strategy to assigning one or more categories, the normalized score k-NN classifier (NSML k-NN) is introduced. The score for each category is divided by the maximum score. The normalized score of the first category in a new ranked list is 1. Moreover, two strategies for calculating scores are introduced, i.e., the maximal of similarity (equation 3) per category and the average of similarity per category (equation 4). The drawback of the strategy from equation 2 is that the high score can be accumulated from many small scores of a set of k nearest neighbors. If this high score of category is assigned, we will belief in many but not brilliant experts. On the other hand, the equation 3 belief in the smartest expert for each category from a set of categories in k nearest neighbors. This is suitable in a training set which the numbers of samples in categories are small but each sample is reliable to use as a training sample. If we believe in a set of brilliant experts which is an advantage of k-NN algorithm, the equation 4 can be applied. The categories belong to the formula, selected when the normalized scores of the categories are greater than or equal to the cutoff point. For example, a list of normalized scores for a formula (Formula01) is shown in Figiure 1.

In Figure 1, ranks of therapeutic categories are shown in the format of "Therapeutic Category:Score". The upper line is absolute scores and the lower line is normalized scores. When the k is 10, the cutoff point is 0.20 (20%) and the strategy of the average score per category, the predicted therapeutic categories of the Formula01 are 02GI and 07Musculo with the normalized scores of 1.00 and 0.64. The third category (01CVS) and the other categories are ignored due to the fact that their normalized score value are less than 0.20 (0.15 for the third category).

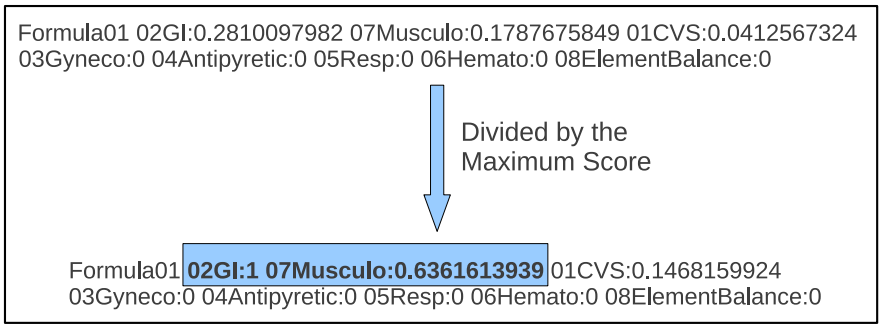


Fig. 1. A set of therapeutic categories is selected by normalized score

5 Experimental Settings and Evaluation

The experimental settings for multi-label herbal formulae and evaluation on experiments are described in this section.

5.1 Experimental Settings of Herbal Formulae

To evaluate the concept of applying text categorization concept to classify multi-label herbal formulae, a set of 58 composite herbal formulae from the National List Essential Medicine 2011 and 230 single herbal formulae from a list of common household remedies for traditional medicines 1999, was used as a training set. The total number of herbal formulae was 288. The minimum number of therapeutic categories was 1 on both types of herbal formulae. The maximum numbers of therapeutic categories were 2 and 5 on composite and single herbal formulae, respectively. The distribution of herbal formulae on each therapeutic categories is shown it Table 1. The category-id (CAT ID), therapeutic categories' name (CAT Name) and the numbers of herbal formulae on each therapeutic category (# Formulae) are presented. The numbers of formulae for 1, 2, 3, 4 and 5 categories are 244, 37, 6, 0 and 1, respectively. Note that the maximum categories of 3 can cover 287 of 288 formulae.

A unigram model was applied in all experiments. The train formula and test formula vectors were normalized by their length. The cosine similarity was used. The value of score was in range from 0 to 1.

5.2 Evaluation

To assign a suitable category to a test document, we can apply the score between the class vector and the test document vector. For finding the best classifier on each dimension, it can be evaluated using measures similar to the traditional

Table 1. The Distribution of the Numbers of Herbal Formulae on Each Category

CAT ID	CAT Name	# Formulae
01	Cardiovascular	38
02	Gastroinstestinal	96
03	Gynecologic	43
04	Antipyretic	85
05	Respiratory	20
06	Blood Tonic	12
07	Musculoskelatal	30
08	Elementary Balance	18

measures for evaluating a ranking-based retrieval system called precision(P), recall (R) and F_1 [4]. Two types of F_1 are used as performance indices, i.e., per-class effectiveness and all-class effectiveness. The per-class effectiveness of a classifier, is the average F_1 from ten trials for each class. The all-class performance of a classifier is calculated by averaging the local measures on a data set. Two types of all-class measures are usually used. For the macro-average, an equal weight is given to the performance on every class, regardless of how large the class is. When the size of each class is also affected to the overall performance, the micro-average is applied. In this work, both the macro-average and micro-average of the F_1 were used for classification of herbal formulae.

6 Experimental Results

Three experiments were conducted to evaluate the proposed method. In the first experiment, two term weighting schemes were used to constructed prototype vectors. The better performance one was selected in the next experiment. The value of maximum category and the cutoff point were set on the selected k-NN classifier from the first experiment. In the last experiment, a set of commercial herbal medicinal products was used to evaluate the proposed method.

6.1 Finding the Appropriated Term Weight of k-NN Classifiers

In order to evaluate the good term weight, three term weighting schemes were used, i.e., tf (W1), $tf \times idf$ (W2), and $(0.5 + 0.5\frac{tf}{max\ tf}) \times idf$ (W3), on a data set of multi-label herbal formulae. The same term weighting scheme is applied on both training and test formulae. The experiment was performed using all formulae as a training set and a test set (test on training set). Furthermore, the values of k were set to 2, 5 and 10. The results of F_1, macro-average F_1 and micro-average F_1 on each class and on each classifier are shown in Table 2.

From the results, some observations could be found. With the default decision of the multiclass classifiers of traditional k-NN, the highest score is used. Only single category is assigned. The W2 and W3 performed better than the W1 when

Table 2. Performance of Text Classifiers by F1 on Therapeutic Classes

CAT ID	CAT Name	F1								
		k=2			*k=5*			*k=10*		
		W1	**W2**	**W3**	**W1**	**W2**	**W3**	**W1**	**W2**	**W3**
01	Cardiovascular	0.95	0.95	0.95	0.42	0.81	0.86	0.25	0.82	0.81
02	Gastroinstestinal	1.00	1.00	1.00	0.83	0.87	0.87	0.36	0.87	0.86
03	Gynecologic	0.99	0.99	0.99	0.66	0.86	0.86	0.27	0.83	0.83
04	Antipyretic	0.98	0.98	0.98	0.67	0.93	0.94	0.40	0.93	0.93
05	Respiratory	1.00	1.00	1.00	0.67	0.85	0.82	0.57	0.84	0.84
06	Blood Tonic	0.96	0.96	0.96	0.59	0.80	0.86	0.15	0.74	0.67
07	Musculoskelatal	1.00	1.00	1.00	0.67	0.93	0.95	0.50	0.93	0.91
08	Elementary Balance	0.97	0.97	0.97	0.36	0.88	0.94	0.00	0.88	0.88
Macro Average F1		**0.98**	**0.98**	**0.98**	0.61	0.87	0.89	0.31	0.85	0.84
Micro Average F1		**0.99**	**0.99**	**0.99**	0.67	0.88	0.89	0.34	0.87	0.86

the k *were* 5 and 10. The performance of the three term weighting schemes is equal and the best when the k was 2. The W2 was selected to perform in the next step.

6.2 Adjusting the Cutoff Point

In this experiment, the cutoff point was set for the W2. The normalized score was applied in this experiment. The cutoff points were set to 20%, 40% and 100% (no cutoff point). The cutoff points at 10% mean we accept the therapeutic categories which their normalized scores $\geq$0.10. The values of k were set to 2, 5 and 10. The results of F_1 on each class, macro-average F_1 and micro-average F_1 and on each value of k and cutoff point, are shown in Table 3.

Table 3. The Results of Classifiers by Using the Cutoff Point

CAT ID	CAT Name	F1								
		k=2			*k=5*			*k=10*		
		20	**40**	**100**	**20**	**40**	**100**	**20**	**40**	**100**
01	Cardiovascular	0.95	0.95	0.95	1.00	0.97	0.81	1.00	0.99	0.82
02	Gastroinstestinal	1.00	1.00	1.00	1.00	0.99	0.87	1.00	0.98	0.87
03	Gynecologic	0.99	0.99	0.99	1.00	0.99	0.86	1.00	0.99	0.83
04	Antipyretic	0.98	0.98	0.98	1.00	0.98	0.93	1.00	0.99	0.93
05	Respiratory	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	0.84
06	Blood Tonic	0.96	0.96	0.96	1.00	1.00	0.80	1.00	1.00	0.74
07	Musculoskelatal	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.98	0.93
08	Elementary Balance	0.97	0.97	0.97	1.00	1.00	0.88	1.00	1.00	0.88
Macro Average F1		0.98	0.98	0.98	**1.00**	0.99	0.87	**1.00**	0.99	0.85
Micro Average F1		0.99	0.99	0.99	**1.00**	0.99	0.88	**1.00**	0.99	0.87

From the result, some conclusions could be described. Although the k=2 was the best in the first experiment, it could not reach the maximum performance due to the nearest neighbors were not enough. Performance of classifiers with values of k = 5 and 10, is better. When the cutoff point was lower, performance of classifiers was better. However, it may lead to over predict for a formula, i.e., a formula may be assigned more categories it should be.

6.3 Evaluation on the Commercial Products

In this experiment, the twelve well-known commercial products in herbal medicine were used as a set of test formulae (F01, F02, ..., F12). The ten formulae had more than one therapeutic category. However, the eight therapeutic categories were taken into account. In order to evaluate the effectiveness of the algorithm, The k-NN classifiers with term weighting schemes of W2 and W3 were investigated The values of k were set to 5 and 10. The cutoff point is set to 20%. Three strategies for calculating the scores were explored, i.e., the maximal sum score (SC1), max score (SC2) and mean score (SC3). The result of F1s is shown in Table 4 and the detail result for each formula when term weighting scheme is W3 and the strategy of SC3, is shown in Table 5. Note that the result of a classifier when the value of k=20, is also reported. The bold category id means the actual category of each formula. The underline category id of k=10 and k=20 means additional predicted category to the predicted categories when the values of k=5 and k=10, respectively.

Table 4. The Results of Classifiers on the Commercial Products

CAT ID	CAT Name	F1											
		k=5						*k=10*					
		W2			W3			W2			W3		
		SC1	SC2	SC3	SC1	SC2	SC3	SC1	SC2	SC3	SC1	SC2	SC3
01	Cardiovascular	0.55	0.55	0.86	0.67	0.75	0.86	0.60	0.86	1.00	0.86	1.00	1.00
02	Gastroinstestinal	0.44	0.44	0.44	0.50	0.57	0.57	0.60	0.83	0.86	0.83	1.00	1.00
03	Gynecologic	0.86	0.86	0.67	0.67	0.67	0.67	0.86	0.86	0.86	0.86	0.86	0.86
04	Antipyretic	0.86	0.86	0.86	0.86	0.75	0.75	0.86	1.00	1.00	1.00	0.86	0.86
05	Respiratory	1.00	1.00	0.67	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00
06	Blood Tonic	0.67	0.67	0.67	0.67	0.67	0.67	0.67	1.00	1.00	1.00	1.00	1.00
07	Musculoskelatal	0.00	0.00	0.00	0.67	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00
08	Elementary Balance	0.67	0.67	0.50	0.67	0.67	0.67	0.67	1.00	1.00	1.00	1.00	1.00
Macro Average F1		0.63	0.63	0.58	0.71	0.76	0.73	0.66	0.82	0.82	0.82	**0.96**	**0.96**
Micro Average F1		0.61	0.61	0.59	0.67	0.71	0.71	0.66	0.83	0.83	0.83	**0.95**	**0.95**

From the result in Table 4, the strategies of SC2 and SC3 are better than the popular SC1. The detail result in Table 5 suggested that less categories predicted when the k=5 was used. Several actual categories were not assigned. The more actual categories were suggested when the k was set to 10. When the k was set to

Table 5. The Detail of Predicted Categories on Each Commercial Product

Formula ID	Actual CAT ID	Predicted CAT ID		
		k=5	k=10	k=20
F01	**01,02**	**01**,05,08	**01,02**,05,08	**01,02**,04,05,06,08
F02	**01,02**	**01**,08	**01,02**,08	**01,02**,08
F03	**01,04**	**01**,**04**	**01**,**04**,08	**01**,02,**04**,08
F04	**04**	01,02,**04**	01,02,**04**	01,02,**04**,08
F05	**02,04**	**04**	**02,04**	01,**02,04**,07
F06	**02,07**	*01,02*	*01,**02**,05,**07***	*01,**02**,05,**07**,08*
F07	**02,07**	**02,07**	**02,07**	**02,07**
F08	**03,08**	01,02,**03**,**08**	01,02,**03**,**08**	01,02,**03**,06,07,**08**
F09	**03**	**03**	**03**	**03**
F10	**01,03,06**	**01**,04,**06**,08	**01**,02,04,**06**,08	**01**,02,04,**06**,08
F11	**03,06,08**	01,**03**	01,**03**,05,**06**,**08**	01,**03**,05,**06**,**08**
F12	**05**	**05**	02,**05**	01,02,03,**05**,08

20, no additional actual categories were assigned while additional incorrect categories were suggested. The value of k is 10 may be the best in this experiment. Although some actual categories were not assigned, e.g., the category 03 for F10, the reason was that this formula was quite different from formulae for this category in the training set, i.e., some Chinese herbs were used in this formula. Some additional categories were over predicted the actual categories, which provided us interested information. This may be minor therapeutic categories in normal dose. It may be effective when higher dose is applied. On the other hand, it may cause some side effects. For example, the F08 may produce side effects on gastrointestinal tract. This information should be advised to patients who would like to take the medicine.

7 Conclusion and Future Work

In this paper, the normalized score multi-label k-NN, was proposed for multi-label herbal formulae classification. With the concept of text categorization, the k-NN classifiers with several term weight schemes were used. The normalized scores were calculated. The values of k, strategies to assign categories were investigated to adjust the decision for multi-label herbal formulae. The experiment was done using a mixed data set of herbal formulae collected from the Natural List of Essential Medicine and the list of common household remedies for traditional medicine. Moreover, a set of well-known commercial products were used for evaluating the effectiveness of the proposed method. From the results, the normalized score multi-label k-NN was an efficient method to classify a herbal formula into one category or multiple categories. Its performance was depended on the set values of the maximum category and the cutoff point. We plan to find the way to construct more advanced strategies to calculate scores. We left this for our future work.

References

1. Lovell-Smith, H.D.: In defence of ayurvedic medicine. The New Zealand Medical Journal 119, 1–3 (2006)
2. Aziz, Z., Peng, T.N.: Herbal medicines: prevalence and predictors of use among malaysian adults. Complementary Therapies in Medicine 44, 44–50 (2009)
3. Roiger, R., Geatz, M.: Data Mining: A Tutorial Based Primer. Addison-Wesley, Boston (2002)
4. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34, 1–47 (2002)
5. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using em. Machine Learning 39, 103–134 (2000)
6. Duwairi, R., Al-Zubaidi, R.: A hierarchical k-nn classifier for textual data. The International Arab Journal of Information Technology 8, 251–259 (2011)
7. Lertnattee, V., Theeramunkong, T.: Effect of term distributions on centroid-based text categorization. Information Sciences 158, 89–115 (2004)
8. Joachims, T.: Learning to Classify Text using Support Vector Machines. Kluwer Academic Publishers, Dordrecht (2002)
9. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523 (1988)
10. Singhal, A., Salton, G., Buckley, C.: Length normalization in degraded text collections. Technical Report TR95-1507 (1995)
11. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal Data Warehousing and Mining 3, 1–13 (2007)
12. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine Learning 85, 333–359 (2011)
13. Fujino, A., Isozaki, H., Suzuki, J.: Multi-label text categorization with model combination based on f1-score maximization. In: Proceeding of The 3rd International Joint Conference on Natural Language Processing, pp. 823–828 (2008)
14. Hua, L.: Research on multi-classification and multi-label in text categorization. In: Proceeding of International Conference on Intelligent Human-Machine Systems and Cybernetics, pp. 86–89 (2009)
15. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048 (2007)
16. Younes, Z., Abdallah, F., Denœux, T.: An Evidence-Theoretic k-Nearest Neighbor Rule for Multi-label Classification, pp. 297–308 (2009)

Unsupervised Approach to Hindi Music Mood Classification

Braja Gopal Patra[1], Dipankar Das[2], and Sivaji Bandyopadhyay[1]

[1] Dept. of Computer Science & Engineering, Jadavpur University, Kolkata, India
[2] Dept. of Computer Science & Engineering, NIT Meghalaya, India
{brajagopal.cse,dipankar.dipnil2005}@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract. We often choose to listen to a song that suits our mood at that instant because an intimate relationship presents between music and human emotions. Thus, the automatic methods are needed to classify music by moods that have gained a lot of momentum in the recent years. It helps in creating library, searching music and other related application. Several studies on Music Information Retrieval (MIR) have also been carried out in recent decades. In the present task, we have built an unsupervised classifier for Hindi music mood classification using different audio related features like rhythm, timber and intensity. The dataset used in our experiment is manually prepared by five annotators and is composed of 250 Hindi music clips of 30 seconds that consist of five mood clusters. The accuracy achieved for music mood classification on the above data is 48%.

Keywords: Hindi Music, Music Mood Classification, MIR, Mood Taxonomy.

1 Introduction

With the rapid evolution of technology, most of the people enhance their life with several technological stuffs. Nowadays, people are enjoying music at leisure time and the overall collection of music increases day by day. However, people are more interested in creating music library which allows them to access songs in accordance with the music moods rather than their title, artists and or genre. Thus, classifying and retrieving music with respect to emotions or mood has become an emerging research area.

Music, also referred as the "language of emotion" can be categorized in terms of its emotional associations [17]. Music perception is highly intertwined with both emotion and the context [6]. The emotional meaning of the music is subjective and thus, it depends upon many factors including culture [7]. Moreover, the mood category of a song varies depending upon several psychological conditions of the Human Beings. Representations of music mood with the psychology remain an active topic for research.

In this paper, we have used a fuzzy c-means classifier for automatic mood classification of Hindi music. As Hindi is the national language of India, Hindi songs are one

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 62–69, 2013.
© Springer International Publishing Switzerland 2013

of the most popular categories of Indian songs and are present in Hindi cinemas or Bollywood movies. Hindi songs make up 72% of the music sales compared to other language songs in India[1]. Mainly, we have concentrated on the collection of Hindi music data annotated with five mood classes[2]. Then, a computational model has been developed to identify the moods of songs using several high and low level audio features. Finally, we have employed fuzzy c-means clustering algorithm and achieved 48% of accuracy on a data set of 250 songs consisting of five mood clusters.

The rest of the paper is organized in the following manner. Section 2 briefly discusses the related work on different languages like English, Indian and Chinese available to date. The dataset and mood taxonomy used in the experiments are described in Section 3. Section 4 describes the list of features for implementing machine learning algorithm. Brief discussion on fuzzy c-means clustering algorithm is described in Section 5. Section 6 presents the experiments with detailed analysis of results. Finally, conclusions are drawn and future directions are presented in Section 7.

2 Related Work

Music classification has received much attention by the researchers in MIR research in the recent years. In the MIR community, Music Information Retrieval Evaluation eXchange[3](MIREX) is an annual competition on several important music information retrieval tasks since 2004. The music mood classification task was included into MIREX in the year of 2007. Many tasks were presented related to English music classification such as Genre Classification, Mood classification, Artist Identification, Instrument Identification and Music Annotation etc.

Considerable amount of work has been done on the music mood classification based on audio, lyrics, social tags and all together or in a multi modal approach as described in [6, 16, 17]. Many tasks have been carried out on the English music mood classification such as lyrics [13, 14], audio [7, 18] and both [1, 6]. Some of the works in Chinese music have been conducted based on audio [3] and lyrics [16].

In contrast to other languages, only a few works on mood detection in Indian music has been reported to date and most of the work can be seen on the Carnatic Music [20, 21]. The performance of mood classification on Indian Classical Music was done in Koduri [20] and Hampiholi [21]. However, there are few works available in Hindi Music Mood Classification based on audio [22] and lyrics and can be seen in [19]. Velankar and Sahasrabuddhe [8] prepared data for Hindustani classical music mood classification. They have performed several sessions for classifying the three Indian Ragas into 13 mood classes. To the best of our knowledge, the fuzzy c-means clustering has not been used in mood classification tasks except the work described in [11] where the similar algorithm was used for genre classification of English songs into 10 genres.

---

[1] http://en.wikipedia.org/wiki/Music_of_India

[2] The term class and cluster are used interchangeably in this paper.

[3] http://www.musicir.org/mirex/wiki/MIREX_HOME

3 Mood Taxonomy and Data Set

One of the issues closely related with mood classification is to identify the appropriate taxonomy for classification. Ekman [9] has defined six basic emotion classes such as *happy, sad, anger, fear, surprise* and *disgust*, but these classes have been proposed for the image emotion classification as we cannot say a piece of music is *disgust*. Thus, in music psychology, our traditional approach is to describe moods using the adjective like *gloomy, pathetic* and *hopeful* etc. However, there is no standard taxonomy available which is acceptable to all the researchers.

Russel [5] proposed the circumplex model of affect based on the two dimensional model. These two dimensions are denoted as "pleasant-unpleasant" and "arousal-sleep". There are 28 affect words in Russel'scircumplex models and are shown in Figure 1. Later on, Thayer [10] adapted Russel's model using the two dimensional energy-stress model. Different researchers used their own taxonomy, which are the subsets of Russel's taxonomy. For example, Katayose et al. [4] used all the adjectives including *Gloomy, Urbane, Pathetic* and *Serious*. Yang et al., [16] used *Contentment, Depression, Exuberance* and *Anxious/Frantic* as mood taxonomy.

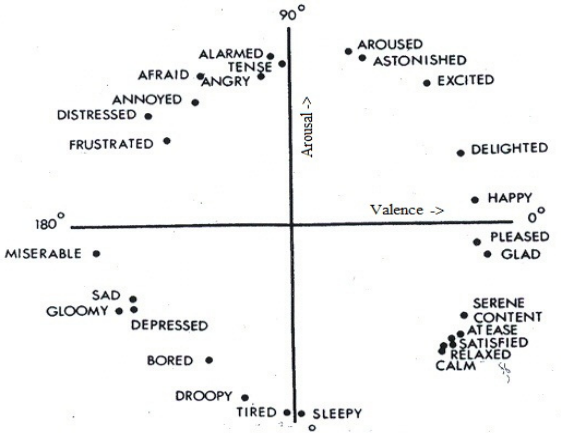


Fig. 1. Russell'scircumplex model of 28 affects words

Table 1. Five mood cluster of proposed mood taxonomy

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Excited	Delighted	Calm	Sad	Alarmed
Astonished	Happy	Relaxed	Gloomy	Tensed
Aroused	Pleased	Satisfied	Depressed	Angry

Our collected data set includes five clusters of moods according to Theyer's model [10] and Russel's Circumplex model [5]. We have also followed the MIREX mood taxonomy [12], which has five mood clusters and each of the clusters has more than

four sub classes. It has been observed that MIREX evaluation forum provides a standard taxonomy for mood classification and many researchers have also used this mood taxonomy [2, 12]. Our mood taxonomy contains five clusters with three subclasses. The mood taxonomy is formed by clustering similar affect words of Russels' circumplex model. For example, we have kept *calm, relaxed* and *satisfied* in one cluster so as to collect similar songs into one group and in this case, the audio features do not vary much. The mood taxonomy used in our experiment is shown in Table 1.

In the present task, a standard data set has been used for the mood classification task. This data has been collected manually and prepared by five human annotators. The songs used in the experiments are collected from Indian Hindi music website[4]. We have faced several problems during the annotation of music. First problem was whether it would be better to ignore the lyrics or not. In Hindi music, we have observed that several songs contain different music as well as different lyrics. For example, a music having high valence consists of the lyric that belongs to the sad mood class. Hu et al. [12] prepared the data based on music only and the lyrics of the song were not considered in their work. So, we also tried to avoid the lyrics of the song as much as possible to build a ground-truth set.

On the other hand, the second problem was the time frame for a song. We have considered only the first 30 seconds of the song so as to prepare our data. In this frame, some lyrics might be present for some of the songs. We have only included the songs that contain lyrics of less than 10 seconds. Finally, we have collected total 250 music tracks out of which 50 tracks were considered from each of the mood clusters. As we have considered only the 30 seconds music from the whole track, it was difficult to identify the track by the annotators. So the inter-annotator agreement was less and was around 72%.

4 Feature Selection

The feature selection always plays an important role in building a good pattern classifier. Thus, we have considered the key features like intensity, timbre and rhythm for our mood classification task. It has been observed that tempo, sound level, spectrum and articulation are highly related to various emotional expressions. Different patterns of the acoustic cues are also associated with different emotional expressions. For example, exuberance is associated with fast tempo, high sound and bright timbre whereas sadness is with slow tempo, low sound and soft timbre.

Rhythm Feature: Rhythm strength, regularity and tempo are closely related with people's moods or responses [3]. For example, generally, it is observed that, in *Exuberance* cluster, the rhythm is usually strong and steady; tempo is fast, whereas in *Depression* cluster, it is usually slow and without any distinct rhythm pattern.

Intensity Feature: Intensity is an essential feature in mood detection and is used in several research works [3, 7]. Intensity of the *Exuberance* cluster is high, and low in

[4] http://www.songspk.name/bollywood_songs.html

Depression cluster. It is observed that the intensity is approximated in general by considering the root mean square (RMS) values of the corresponding signal.

Timbre Feature: Many existing researchers have shown that mel-frequency cepstral coefficients (MFCCs), so called spectral shapes and spectral contrast are the best features for identifying the mood of music [3, 7, 18]. In this paper, we have used both spectral shape and spectral contrast. Spectral shape includes brightness or centroid, band width, roll off and spectral flux. Spectral contrast features includes sub-band peak, sub-band valley, sub-band contrast.

All the features used in our experiments are listed in Table 2. These features are extracted using jAudio[5] toolkit. It is a music feature extraction toolkit developed in JAVA platform. The jAudio toolkit is publicly available for research purpose.

Table 2. Feature used in mood classification

Class	Features
Rhythm	Rhythm strength, regularity and tempo
Timbre	MFCCs, Spectral shape, Spectral contrast
Intensity	RMS energy

5 Fuzzy Clustering

We have built an unsupervised classifier to classify the music files into five clusters as stated above in Section 3. We implemented Fuzzy C-means clustering algorithm for the classification purpose. The membership functions of each music vary in between 0 and 1. A detail of the algorithm is given below.

5.1 Fuzzy C-Means Clustering Algorithm

For unsupervised fuzzy clustering, we have chosen the well-known fuzzy c-means clustering algorithm which was already used in music genre classification task [11].

Let there are N data points i.e., N={x_1, x_2, x_3……x_n} and each data points are represented by p dimensional feature space i.e. x_k= {x_{k1}, x_{k2}, x_{k3} … x_{kp}}.

The main objective of fuzzy c-means clustering algorithm is to classify p dimensional data points N into a set of c fuzzy classes of centroids v_1, v_2, v_3 … v_c in same feature space, such that the sum of membership function of any data point x_k, in all classes is 1 [11].

Membership function can be represented by,

$$\sum_{i=1}^{c} \mu_{ik} = 1, \text{ for k=1 to N.} \quad (1)$$

For all c clusters, we can derive the cluster centers v_i for i=1 to c by given equation.

[5] http://sourceforge.net/projects/jmir/files/

$$v_i = \frac{\sum_{k=1}^{N} \mu_{ik}^{p} x_k}{\sum_{k=1}^{N} \mu_{ik}^{p}} \tag{2}$$

From the above equation, we can see that the cluster center v_i is basically the weighted average of the membership μ_{ik}. To find out the optimal distribution of points of the clusters and optimal placement of centroids, we require an objective function J_p over $\mu = \{\mu_{ik}\}$ and $v = \{v_i\}$, and can be represented as:

$$J_p(\mu, v) = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^{m} \|x_k - v_i\|^2 \tag{3}$$

Where $\| \|$ is the inner product induced norm in p dimension and Here $m>1$ is any real number and it influences the membership grade. Then the membership function is updated by the expression given below:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{\frac{2}{m-1}}} \tag{4}$$

In the above experiment, we have used m=2, which influence the membership grade. To get desired result, we performed the algorithm iteratively:

- Assigning random values to all v_i and μ_{ik} at initial stage.
- Iteratively recalculate the values for all v_I and then all μ_{ik} according to equation (2) and (4).
- Stop, when the objective function J changes from the previous iteration less than by a small number δ, a given parameter (here, we used $\delta = 0.01$).

6 Experiments and Evaluation

In order to achieve good results, we require a huge amount of mood annotated music corpus for applying the clustering algorithm. But, to the best of our knowledge, no mood annotated Hindi songs are available to date. Thus, we have developed the data-set by ourselves and it contains 250 songs consisting of five clusters. We have used the fuzzy c-means clustering algorithm described in Section 5 to accomplish our classification experiments based on the features we discussed in Section 4. The features are extracted using the jAudio Feature Extractor.

The accuracies have been calculated and are reported in Table 3. We have considered that a song belongs to *cluster 1* if membership function of *cluster 1* ($\mu_{cluster1}$) is greater than all other membership functions. The confusion matrix of the classification accuracy is given in Table 4. We have achieved the maximum accuracy of 52% in *cluster 3*. It has been observed that the *cluster 1* and *cluster 5* have lowest accuracy and is about 46%. The accuracies of *cluster 2* and *cluster 4* are same and is 48%.

We have observed that some of the instances from each of the clusters have tendency to go towards its neighboring clusters. For example, some songs from *cluster 2* fall under the *cluster 1* as they have similar RMS energy and tempo. The accuracy achieved by the system is quite low as compared to the other existing mood classification systems for

Table 3. Accuracies of each class

Class	Accuracy
Cluster 1	46
Cluster 2	48
Cluster 3	52
Cluster 4	48
Cluster 5	46
Average	48

Table 4. Confusion matrix for the accuracy

		Predictions				
	Clusters	**1**	**2**	**3**	**4**	**5**
T	**1**	**23**	8	4	3	12
r	**2**	11	**24**	3	4	8
u	**3**	3	10	**26**	8	3
e	**4**	3	3	13	**24**	7
	5	15	3	5	4	**23**

English songs [3, 7], Chinese songs [16], Hindi Songs [19, 22] and Carnatic songs [20, 21]. Later on, inclusion of additional features and the feature engineering may remove such kind of biasness and improve the results.

7 Conclusion and Future Work

In this paper, we have developed the fuzzy c-means classifier for Hindi music mood classification based on the simple audio features namely rhythm, intensity and timbre. We have used our own mood taxonomy described in Section 3 and tried to generalize in accordance to MIREX mood taxonomy and Russel's Circumplex model. We have achieved low accuracy of 48%.

There are several directions for future work. One of them is to improve the fuzzy c-means clustering algorithm using modified objective function as stated in [11]. Incorporate more audio features for enhancing the current results of mood classification. Later on lyrics of the song may be included for multimodal mood classification. Preparing the large audio data and collecting the lyrics of those songs may be considered as the other future direction of research.

References

1. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Proceedings of the Seventh International Conference on Machine Learning and Applications (ICMLA 2008), pp. 688–693. IEEE (2008)
2. Laurier, C., Sordo, M., Herrera, P.: Mood cloud 2.0: Music mood browsing based on social networks. In: Proceedings of the 10th International Society for Music Information Conference (ISMIR 2009), Kobe, Japan (2009)
3. Liu, D., Lu, L., Zhang, H.J.: Automatic Mood Detection from Acoustic Music Data. In: Proceedings of the International Society for Music Information Retrieval Conference, ISMIR 2003 (2003)
4. Katayose, H., Imai, H., Inokuchi, S.: Sentiment extraction in music. In: Proceedings of the 9th International Conference on Pattern Recognition, pp. 1083–1087. IEEE (1988)
5. Russell, J.A.: A Circumplx Model of Affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
6. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music Mood and Theme Classification-a Hybrid Approach. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 657–662 (2009)

7. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech, and Language Processing 14(1), 5–18 (2006)
8. Velankar, M.R., Sahasrabuddhe, H.V.: A Pilot Study of Hindustani Music Sentiments. In: Proceedings of 2nd Workshop on Sentiment Analysis Where AI Meets Psychology (COLING 2012), IIT Bombay, Mumbai, India, pp. 91–98 (2012)
9. Ekman, P.: Facial expression and emotion. American Psychologist 48(4), 384–392 (1993)
10. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford University Press, Oxford (1989)
11. Poria, S., Gelbukh, A., Hussain, A., Bandyopadhyay, S., Howard, N.: Music Genre Classification: A Semi-supervised Approach. In: Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Rodríguez, J.S., di Baja, G.S. (eds.) MCPR 2012. LNCS, vol. 7914, pp. 254–263. Springer, Heidelberg (2013)
12. Hu, X., Downie, S.J., Laurier, C., Bay, M., Ehmann, A.F.: The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2008), pp. 462–467 (2008)
13. Hu, X., Downie, S.J., Ehmann, A.F.: Lyric text mining in music mood classification. In: Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 411–416 (2009)
14. Hu, Y., Chen, X., Yang, D.: Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), pp. 123–128 (2009)
15. Yang, Y.H., Liu, C.C., Chen, H.H.: Music emotion classification: a fuzzy approach. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 81–84. ACM (2006)
16. Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., Chen, H.H.: Toward multimodal music emotion classification. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 70–79. Springer, Heidelberg (2008)
17. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D.: Music emotion recognition: A state of the art review. In: Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR 2010), pp. 255–266 (2010)
18. Fu, Z., Lu, G., Ting, K.M., Zhang, Z.: A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia 13(2), 303–319 (2011)
19. Ujlambkar, A.M., Attar, V.Z.: Mood classification of Indian popular music. In: Proceedings of the CUBE International Information Technology Conference, pp. 278–283. ACM (2012)
20. Koduri, G.K., Indurkhya, B.: A Behavioral Study of Emotions in South Indian Classical Music and its Implications in Music Recommendation Systems. In: Proceedings of the 2010 ACM Workshop on Social, Adaptive and Personalized Multimedia Interaction and Access, pp. 55–60. ACM (2010)
21. Hampiholi, V.: A method for Music Classification based on Perceived Mood Detection for Indian Bollywood Music. World Academy of Science, Engineering and Technology 72, 507–514 (2012)
22. Patra, B.G., Das, D., Bandyopadhyay, S.: Automatic Music Mood Classification of Hindi Songs. In: Proceedings of 3rd Workshop on Sentiment Analysis where AI meets Psychology (IJCNLP 2013), Nagoya, Japan, pp. 24–28 (2013)

Motion Intensity Code for Action Recognition in Video Using PCA and SVM

J. Arunnehru[1] and M. Kalaiselvi Geetha[2]

[1,2] Department of Computer Science and Engineering, Annamalai University, India
{arunnehru.aucse,geesiv}@gmail.com

Abstract. Manual video surveillance is highly expensive and inconvenient in continuous monitoring, by a security personnel. So automatic video surveillance and activity recognition is needed. In this paper, an activity recognition approach is proposed, the difference image is used to extract the motion information based on Region of Interest (ROI). The experiments are carried out on KTH dataset, considering four activities viz (walking, running, waving and boxing) and Weizmann dataset, considering four activities viz (walking, running, waving one hand, waving both hands) with Support Vector Machines (SVM) for classification. This approach shows an overall performance of 94.75% using KTH dataset and 92% using Weizmann dataset to recognize the actions. The performance of the proposed approach is comparable with well known existing methods.

Keywords: Video Surveillance, Activity Recognition, Gesture Recognition, Principal Component Analysis, Support Vector Machines, Motion Intensity Code.

1 Introduction

Video surveillance is attracting much of the researcher's attention since it is an indispensable tool for protecting people, public property and finds most promising applications in computer vision. Video surveillance systems are needed for a well-organized surveillance of unusual human activity in public areas like airports, banks, railway stations, ATM, bus stands and commercial buildings. It is very useful for law enforcement to uphold public control and prevent the criminal activity. The aim of the activity recognition is to classify the current events from video. Each human activity is identified by sensuous observations of the movement in body structure, which shows the semantic meaning of the activities like walking, running, waving, bending, boxing and etc., Recognizing the human actions automatically is challenging task due to various factors like clothing, illumination changes, speed variations, changing environments and occlusion. Human body shows large fluctuations in size, appearance and shape, while same action is performed by different individuals.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 70–81, 2013.
© Springer International Publishing Switzerland 2013

1.1 Related Work

Recent surveys in the area of human action analysis in [1] and [2] focus on the feature descriptor, representation and classification model in video sequences. Survey by P. Turaga et. al. [3] centers around recognition of human activity. D. Weinland et. al. [4] gives a general overview and classification of the approaches used in the research. "Bag of-words" representations for object recognition problems in computer vision is gaining success in the computer vision field. Y. Wang. et. al. [5] assigns visual word to each frame of an image sequence by analyzing the motion of the person in the frame. The human silhouettes were used as action description and temporal templates called Motion History Image (MHI) and Motion Energy Image (MEI) are used to represent the actions [6] and matching was done by Hu moments or SVM classifier [7]. L. Wang et. al. [8] presents a frame based learning discriminative feature method for action recognition. Human shape characteristics and motion are identified by using optical flow and edge features, where these two features are combined to recognize the action in video sequences. A novel method for human action recognition based on key frame selection and Pyramid Motion Features (PMF) from a video sequence, which contain the optical flow and biologically inspired features are used to extract discriminate information from each frame of the action sequences. The motion information is represented by key frames, where these key frames are selected by adaboost learning algorithm [9] and SVM classifier are used to recognize the actions.

1.2 Outline of the Work

This paper deals with activity recognition that aims to understand human actions from video sequences. The proposed method is evaluated using KTH [10] and Weizmann [11] action dataset with the person showing actions such as walking, running, waving one hand, waving both hands and boxing. Difference image is obtained by subtracting the successive frames. Motion information is extracted by identifying the Region of Interest (ROI). The extracted ROI is divided into three blocks B1, B2 and B3. Motion Intensity Code (MIC) is extracted as feature from the block showing maximum motion in the ROI. The extracted feature is fed to the SVM classifier for activity recognition.

The rest of the paper is organized as follows. Principal Components Analysis (PCA) is given in Section 2. Section 3 presents an introduction on Support Vector Machine (SVM). Section 4 describes the feature extraction process and related discussion. Section 5 explains the workflow of the proposed approach. Section 6 presents the Experimental results and Section 7 concludes the paper.

2 Principal Component Analysis

PCA is a useful statistical procedure that has found importance in many fields, and is a well-known technique for finding patterns in data of high-ceilinged

dimension. PCA *'combines'* the essence of attributes by creating an alternative, smaller set of variables [12,13]. The initial data can then be projected onto this smaller set.

Suppose that $x_1, x_2, ..., x_p$ are P training vectors, each belonging to one of N classes $\{\zeta_1, \zeta_2, ..., \zeta_N\}$. Then, the training vector, x_p, can be projected to lower dimension vector y_p, using an orthonormal linear transform given by $y_p = W^T x_p$. The transformation matrix ($\mathbf{W}$) can be obtained from the eigenvalues and eigenvectors of the covariance matrix ($\sum$) of the input data. By definition, the covariance of the input data can be estimated as

$$\sum = \frac{1}{P} \sum_{p=1}^{P} (x_p - \mu)(x_p - \mu)^T, \tag{1}$$

Where μ is the mean vector of all the training images.

The eigenvectors of the covariance matrix are $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K$ associated with eigen values $\lambda_1 \geq \lambda_2 \geq ... \lambda_K$ respectively, where K is the feature vector dimension. The transformation matrix ($\mathbf{W}$) can be obtained by retaining $D(D \ll K)$ eigenvectors corresponding to D maximum eigenvalues, i.e., $(\mathbf{W}) = [\mathbf{e}_1\ \mathbf{e}_2 ... \mathbf{e}_D]$. Since is the spread (variance) of the feature population along the direction ($\mathbf{e}_i$), and the amount of information in a population increases with the spread, feature $[\mathbf{e}_1\ \mathbf{e}_2 ... \mathbf{e}_D]$ retain, for any D, the significant part of information in the feature population.

A given test data is also projected to the lower dimension space (say vector $\mathbf{t}$). Then, by minimum distance matching, the test data can be assigned to the class corresponding to the training feature vector $\mathbf{x}_{i_o}$, where $i_o = \arg \min_{1 \leq i \leq p} \|\mathbf{t} - \mathbf{y}_i\|$, where $\| \ . \ \|$ represents the Euclidean distance in $\mathbb{R}^D$. The first eigenvector corresponds to the direction of maximum variance of the zero mean two dimensional data. The second eigenvector is orthogonal to the first eigenvector and it corresponds to the direction of next maximum variance.

3 Support Vector Machines

Support Vector Machine (SVM) is a popular technique for classification in visual pattern recognition [14,15]. The SVM is most widely used in kernel learning algorithm. It achieves reasonably vital pattern recognition performance in optimization theory [16,17]. A classification task are typically involved with training and testing data. The training data are separated by $(x_1, y_1), (x_2, y_2), ...(x_m, y_m)$ into two classes, where $x_i \in \mathbb{R}_N$ contains n-dimensional feature vector and $y_i \in \{+1, -1\}$ are the class labels. The aim of SVM is to generate a model which predicts the target value from testing set. In binary classification the hyper plane $w.x + b = 0$, where $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ is used to separate the two classes in some space $\mathbb{Z}$ [18]. The maximum margin is given by $M = 2/||w||$. The minimization

problem is solved by using Lagrange multipliers $\alpha_i (i = 1, ...m)$ where w and b are optimal values obtained from Eq. 2.

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i \ K(x_i, x) + b\right) \tag{2}$$

The non-negative slack variables ξ_i are used to maximize margin and minimize the training error. The soft margin classifier obtained by optimizing the Eq. 3 and Eq. 4.

$$\min_{w,b,\xi} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{3}$$

$$y_i(w^T \phi(x_i) + b \geq 1 - \xi_i, \xi_i \geq 0 \tag{4}$$

If the training data is not linearly separable, the input space mapped into high dimensional space with kernel function $K(x_i, x_j) = \phi(x_i).\phi(x_j)$ is explained in [17].

4 Motion Intensity Code (MIC) for Action Recognition

Feature is a descriptive characteristic extracted from an image or video sequences, which represent the meaningful data that is vital for further analysis. The following subsections present the description of the feature used in this work.

4.1 Frame Differencing

Motion is a most important dynamic information used to recognize the human activity. The difference image obtained by simply subtracting the current frame is at time $t + 1$ with previous frame t on a pixel by pixel basis. The extracted motion information is considered as the Region of Interest (ROI). Fig. 1(a), Fig. 1(b) shows the two successive frames of the Weizmann dataset. Resulting difference image is show in Fig. 1(c).

(a) (b) (c)

Fig. 1. (a), (b) Two Consecutive frames. (c) Difference image of (a) and (b) from Weizmann dataset.

$$D_k(i,j) = |I_k(i,j) - I_{k+1}(i,j)| \\ 1 \leq i \leq w, 1 \leq j \leq h \tag{5}$$

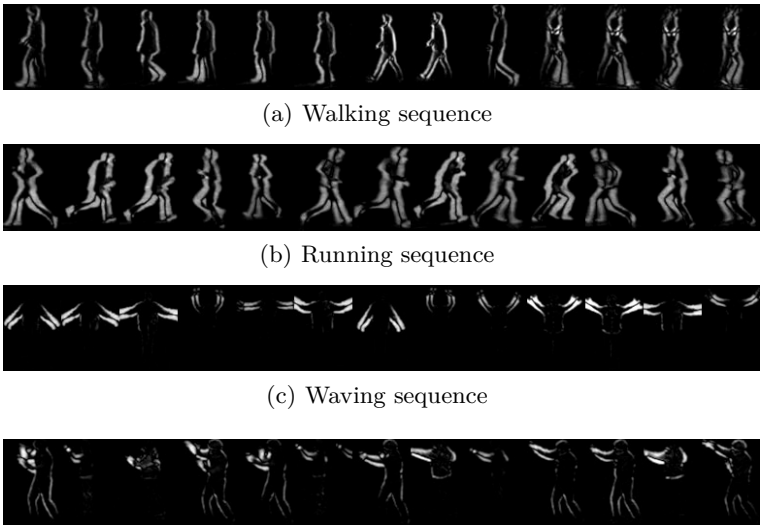
$D_k(i,j)$ is the difference image, $I_k(i,j)$ is the intensity of the pixel (i,j) in the k^{th} frame, w and h are the width and height of the image respectively. Motion information T_k or difference image is calculated using

$$T_k(i,j) = \begin{cases} 1, & \text{if } D_k(i,j) > t; \\ 0, & \text{otherwise;} \end{cases} \tag{6}$$

Where t is the threshold.

4.2 Motion Intensity Code (MIC)

To recognize the action performed, motion is an important cue normally extracted from video. In that aspect, Motion Intensity Code (MIC) is extracted from the motion information extracted from the video sequences. The procedure for extracting the feature is explained in this section.



(a) Walking sequence

(b) Running sequence

(c) Waving sequence

(d) Boxing sequence

Fig. 2. Extracted ROI from KTH dataset

Fig. 2 shows the motion information extracted for different actions from KTH dataset. Initially motion identified region is considered as ROI, as seen in Fig. 3(a). ROI divided into three blocks B1, B2, B3 comprising of head, torso and leg regions as shown in Fig. 3(b). In order to minimize the amount of calculation, only the maximum motion identified block is considered for further analysis. In the Fig. 3(b), block B3 shows maximum motion. So, B3 alone

is considered for MIC extraction as shown in Fig. 3(c). The block under consideration is of size MxN, where $M = 30$ and $N = 60$. This MxN region is divided into k x j subblocks, each of size M/k x N/j, where $k = 2, 3, 6, 10, 15$ and $j = 4, 6, 12, 20, 30$. The values of k and j are fixed in such a way that each subblocks are of equal size. For example, if $k = 3$ and $j = 6$, each subblock will be of size 10x10 pixels as seen in Fig. 3(d). Accordingly for the respective values of k and j, $n-$dimensional feature vectors are extracted from the ROIs and it is fed to the multiclass SVM for further processing.

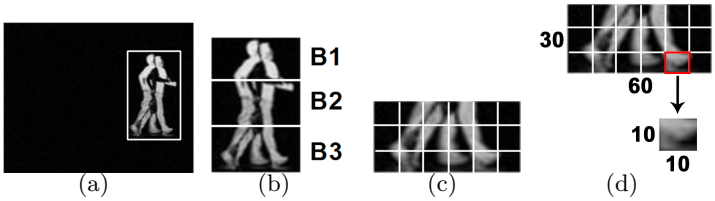


Fig. 3. (a) Motion information. (b) ROI extracted from (a). (c) Block description of (b). (d) 3x6 subblocks

5 Workflow of the Proposed MIC Action Recognition Algorithm

KTH and Weizmann action datasets are used for experimental purpose. The video is processed at 25 frames per second. Smoothing is done by Gaussian convolution with a kernel of size 3 x 3 and variance $\sigma = 0.5$. It is essential to preprocess all video sequences to remove noise for fine features extraction and classification. ROI is extracted from the video sequence and MIC features are extracted as discussed in Section 4. Dimensionality reduction is done with PCA and projected features are fed to the SVM classifier for activity recognition. The workflow of the proposed approach is shown in Fig 4.

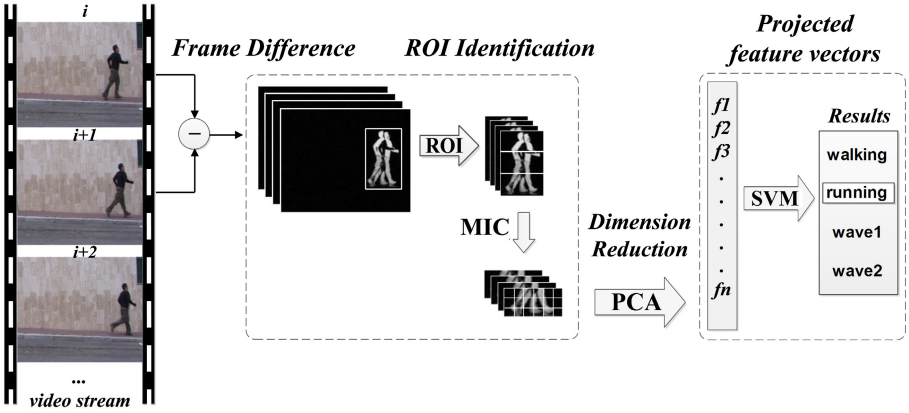


Fig. 4. Workflow of the proposed approach

5.1 MIC Action Recognition Algorithm

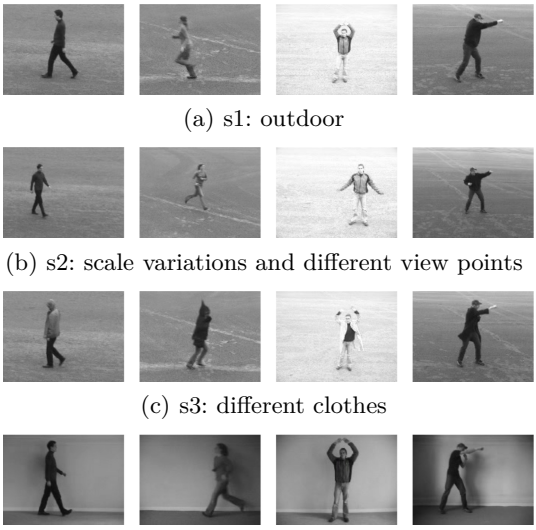
1. Frame differencing is performed with Eq. 5.
2. Motion information is extracted from step 1 using Eq. 6.
3. Motion Intensity Code (MIC) is extracted as explained in Section 4.2.
4. Dimension reduction is performed with PCA and the projected feature vectors are fed to SVM for training.

6 Experimental Results

In this section, proposed method is evaluated using with KTH and Weizmann datasets. The experiments are carried out in C++ with OpenCV 2.2 in Ubuntu 12.04 Operating System on a computer with Intel Xeon Processor 2.40 GHz with 4 GB RAM. The obtained MIC features are fed to LIBSVM [18] tool to develop the model for each activity and these models are used to test the performance. RBF kernel is used for experimental purpose, which non-linearly separates the training data into a higher dimensional space unlike the linear kernel.

6.1 Dataset

KTH Dataset: In KTH, four different actions such as walking, running, waving and boxing are considered for experimental purpose. 25 different persons performed various actions in various scenarios like $s1$: outdoor, $s2$: outdoor with scale variations, $s3$: outdoor with different clothes and $s4$: indoors. Sample frames of action sequence are shown in the Fig. 5.



(a) s1: outdoor

(b) s2: scale variations and different view points

(c) s3: different clothes

(d) s4: indoor with illumination changes

Fig. 5. Sample frames from the KTH dataset

The sequences were taken over static backgrounds. Video data are at 25 fps, each video clip contains one actor performing one action in four different scenarios as explained earlier. Four actions walking, running, waving and boxing taken from 4 different scenarios are used in the experiments. Each clip is of 1 sec duration and for each action, a total of 40 clips are utilized, that includes all four scenarios considered in this work. In this work, 10 persons are taken randomly from four scenarios for evaluation. The samples are divided into a training set of (7 persons), and testing set of (3 persons).

Weizmann Dataset: In Weizmann, four different actions performed by nine different persons, like walking, running, waving one hand and waving both hands are considered for experimental purpose, where each actor performing one action. Video clips are at 25 fps. The sequences were taken over static backgrounds. The sample frames of action sequence are shown in Fig. 6.

Fig. 6. Sample frames from the Weizmann dataset

Four actions like walking, running, wave1 (waving one hand) and wave2 (waving both hands) are used in this experiments. Each clip is of 1 sec duration and for each action, a total of 36 clips are utilized. In this work, nine persons are considered for evaluation. For conducting the experiments, actions performed by seven persons are taken as training samples and the remaining two persons are considered for testing.

6.2 Performance Evaluation

As explained in Section 4, the $n-$dimensional feature vectors are extracted. The projected features are fed to the SVM classifier. To evaluate the performance of the proposed approach, various experiments are carried out on different subblock sizes. In this experiment, the subblock size is empirically fixed for MIC feature extraction. Table 1 and Table 2 shows the accuracy results obtained for different subblock size in KTH and Weizmann dataset respectively. As seen, the subblock size of 3x6 gives better performance. To compute accuracy, $F_\alpha = 2PR/(P + R)$ is utilized, where P, R are Precision and Recall respectively. F-measure (F_α) is the combined measure of accuracy and the weighting factor $\alpha = 1$ is used.

Table 1. On KTH dataset with different subblock size

Subblock	Walking (%)	Running (%)	Waving (%)	Boxing (%)
15 x 30	84	77	85	80
10 x 20	86	78	89	87
6 x 12	92	81	95	89
3 x 6	**97**	**88**	**97**	**97**
2 x 4	89	82	91	90

Table 2. On Weizmann dataset with different subblock size

Subblock	Walking (%)	Running (%)	Wave1 (%)	Wave2 (%)
15 x 30	78	72	86	85
10 x 20	80	76	85	87
6 x 12	85	77	89	90
3 x 6	**90**	**84**	**96**	**98**
2 x 4	78	72	92	94

SVM with RBF Kernel: This approach utilized SVM with polynomial and RBF kernels. The results obtained with RBF kernel are found be satisfactory. While analyzing the performance of the RBF kernel for the MIC features, a grid search has been carried out to identify the best parameters for RBF kernel in parameter space using LIBSVM. In KTH dataset the parameters are $C = 32$, $\gamma = 2.0$ and for Weizmann dataset the parameters are $C = 512$, $\gamma = 2.0$, where C is the slack variable or error weight and γ is the curvature of the decision boundary, where these two parameters is used to obtain the optimal classifier performance. Table 3 and Table 4 shows the classification results of the SVM classifier with RBF kernel for the 3x6 subblock size. Four actions are considered from KTH (walking, running, waving and boxing) and Weizmann (walking, running, waving one hand and waving two hands) datasets, where correct responses define the main diagonal, the majority of actions are correctly classified. As seen in Table 4, since the running and walking actions are similar, some of the running sequences are misclassified as walking and vice versa. Thus, it needs further attention. The average performance of the proposed method for Weizmann dataset is 92%, which is given in Table 4. The average performance of the proposed method on KTH dataset is 94.75%.

Table 3. Confusion matrix for KTH datset using SVM with RBF kernel **(94.75%)**

	Walking (%)	Running (%)	Waving (%)	Boxing (%)
Walking	**0.97**	0.03	0.00	0.00
Running	0.12	**0.88**	0.00	0.00
Waving	0.00	0.02	**0.97**	0.01
Boxing	0.00	0.00	0.03	**0.97**

Table 4. Confusion matrix for Weizmann datset using SVM with RBF kernel **(92%)**

	Walking (%)	Running (%)	Wave1 (%)	Wave2 (%)
Walking	**0.90**	0.10	0.00	0.00
Running	0.14	**0.84**	0.02	0.00
Wave1	0.02	0.00	**0.96**	0.02
Wave2	0.00	0.00	0.02	**0.98**

Accuracy Obtained with Different Subblock Size: The overall performance of the proposed MIC with different subblock size on KTH and Weizmann dataset is shown in Fig. 7. The best performance is achieved with subblock size 3x6 in both the datasets. The classification accuracy increases when the subblock size is increased.

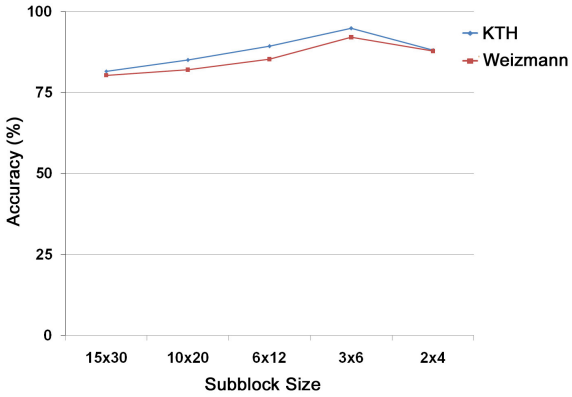


Fig. 7. The evaluation of choosing different subblock sizes for MIC on KTH and Weizmann dataset

Comparsion: Table 5 and Table 6 compare the activity recognition results of the proposed approach with some of the state of the art methods for KTH and Weizmann datasets respectively. Based on the comparison, it is seen that the proposed method shows good results on both KTH and Weizmann action datasets.

Table 5. Comparsion with various methods on KTH dataset

Method	Accuracy(%)
Proposed Approach	**94.75**
Jhuang et al. [19]	91.70
Nowozin et al. [20]	87.04
Dollar et al. [21]	81.17
Schuldt et al. [10]	71.72

Table 6. Comparsion with various methods on Weizmann dataset

Method	Accuracy(%)
Proposed Approach	**92**
Jia *et al.* [22]	90.9
Liu *et al.* [23]	89.3
Thurau [24]	86.7
Klaser *et al.* [25]	84.3

7 Conclusion and Future Work

This paper presented a method for activity recognition for video surveillance using Motion Intensity Code (MIC) as features. Experiments are conducted on both datasets, KTH dataset considering different actions viz (walking, running, waving and boxing) and Weizmann dataset considering actions viz (walking, running, waving one hand and waving two hands). The ROI extracted from the difference image are used for classification based on motion information. This approach then evaluates the performance of motion feature in video sequence using multiclass SVM with RBF kernels. The system gives a good classification of accuracy of 94.75% for KTH dataset and 92% for Weizmann dataset. It is observed from the experiments that the system could not distinguish running and walking with high accuracy and is of future interest.

Acknowledgments. The authors greatfully acknowledge **University Grants Commission of India** [F. No. 41-636/2012 (SR)], for funding this work.

References

1. Poppe, R.: Vision-based human motion analysis: an overview. Computer Vision and Image Understanding (CVIU) 108(1-2), 4–18 (2007)
2. Poppe, R.: A survey on vision-based human action recognition. IVC 28, 976–990 (2010)
3. Turaga, P., Chellappa, R., Venkatramana Subrahmanian, S., Octavian Udrea, O.: Machine recognition of human activities: a survey. IEEE Transactions on Circuits and Systems for Video Technology 18(11), 1473–1488 (2008)
4. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding (CVIU) 115(2), 224–241 (2011)
5. Wang, Y., Mori, G.: Human Action Recognition by Semilatent Topic Models. PAMI 31, 1762–1774 (2009)
6. Bobick, F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
7. Meng, H., Pears, N., Bailey, C.: A human action recognition system for embedded computer vision application. Computer Visual. Pattern Recognition, 1–6 (2007)

8. Wang, L., Wang, Y., Jiang, T., Zhao, D., Gao, W.: Learning discriminative features for fast frame based action recognition. Pattern Recognition 46(7), 1832–1840 (2013)
9. Liu, L., Shoa, L., Rockett, P.: Boosted key-frame selection and correlated Pyramid motion-feature representation for human action recognition. Pattern Recognition 46, 1810–1818 (2013)
10. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: International Conference on Pattern Recognition, vol. 3, pp. 32–36 (2004)
11. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Proc. IEEE International Conferences on Computer Vision, pp. 1395–1402 (2005)
12. Pearson, K.: On lines and planes of closest fit to systems of points in space. Phil. Mag. 2, 559–572 (1901)
13. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Phil. Mag. 24, 417–441 (1933)
14. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)
15. Mitchell, T.: Machine Learning. McGraw-Hill Computer science series (1997)
16. Vapnik, V.: Statistical Learning Theory. Wiley, NY (1998)
17. Lewis, J.P.: Tutorial on SVM. CGIT Lab, USC (2004)
18. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27 (2011)
19. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition, In Proc. of ICCV (2007)
20. Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: Proc. of ICCV (2007)
21. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 14th International Conference on Computer Communications and Networks, pp. 65–72 (2005)
22. Jia, K., Yeung, D.Y.: Human action recognition using local spatio-temporal discriminant embedding. In: Proc. CVPR, pp. 1–8 (2008)
23. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: Proc. CVPR, pp. 1–8 (2007)
24. Thurau, C.: Behaviour histograms for action recognition and human detection. In: International Workshop on Human Motion with ICCV (2007)
25. Klaser, A., Marszalek, M., Schmid, C.: A Spatio-temporal descriptor based on 3D-gradients. In: Proc. BMVC, pp. 1–8 (2007)

Performance Analysis of Tree Based Classification Algorithms for Intrusion Detection System

G.V. Nadiammai and M. Hemalatha

Dept. of Computer Science, Karpagam University, Coimbatore - 21
{gvnadisri,csresearchema}@gmail.com

Abstract. Intruders attack both commercial and corporate distributed systems successfully. The problem of intruders has become vital. The most effective resistance today is the use of Intrusion Detection Systems. An intrusion detection system analysis all aspects of network activities in order to identify the existence of unusual patterns that may represent a network or system attack made by intruders attempting to compromise a system. This paper brings an idea of applying data mining algorithms to the intrusion detection system. Performance of various tree based classifiers like Decision Stump, BF Tree, ID3, J48, LAD, Random Tree, REP Tree, Random Forest and Simple Cart algorithms are compared and the experimental study shows that the Random Forest algorithm outperforms than other algorithms in terms of accuracy, specificity and sensitivity and Time.

Keywords: Data Mining, Intrusion Detection, Machine Learning, Tree based Classifiers, KDD Cup Dataset.

1 Introduction

Security is a major issue in internet and IT industry. Security can be deployed using the technologies like ID, encryption, firewall and access control. But these technologies inhibits possible flaws. The patterns of user activities and audit logs must be analyzed to promote security. An Intrusion Detection System is defined as the process of monitoring the activities that take place over computer, network that differs from the usual behavior of the system. If not, it ignores otherwise it raises alarms to make the administrator to handle the situation. So IDS serves as a gateway for providing secure network. To make IDS effective and to develop the accuracy of intrusion detection, data mining concept was introduced.

Data mining analyzes useful information from large volumes of database that are noisy, fuzzy & random [1]. IDS includes two types of detection approaches like a) Misuse/Signature detection- It accurately identifies the known attacks with less false alarm rate but fails to detect the unknown attacks. The detection efficiency of this method is quite high. b) Anomaly Detection- Efficiently detects the new sorts of attacks but with a high false alarm rate. Two different types of IDS are Host based and Network based IDS. In HIDS the intrusion detection in a single system. It monitors the application program, audit logs of the particular host and compared it with the IDS to check whether any intrusion happens or not. It comes under passive

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 82–89, 2013.
© Springer International Publishing Switzerland 2013

component. In NIDS centrally one IDS is connected to whole LAN. It is a kind of active component and detects denial of service attacks, port scans and network traffic attacks successfully. Classification is a tree based structure involve in the concept of data mining technique, used to predict data instances through attributes. It is used for classifying data into a known set of classes. This approach mainly used to build a model using a training data set and validate it by test data. Moreover the detection rate of the supervised classification will be much better than unsupervised classification depending upon the prior knowledge. In particular, nine Tree based classifiers such as Decision Stump, BF Tree, ID3, J48, LAD, Random Tree, REP Tree, Random Forest and Simple Cart have been utilized to evaluate the classification accuracy.

The paper is organized as follows. Section 2 describes the related work of the study. In section 3 the performance evaluation metrics are explained in detail. The section 4 includes the methodology of the study. Section 5 explains the experimental result and analysis. Conclusion and future work are shown in section 6.

2 Related Work

Robert E. Banfield [2] has evaluated bagging and randomization based approaches in order to form an ensemble of decision tree classifiers for involving possible number of classes. Experiments were made on publicly available datasets. The results of cross validation are more accurate than bagging concepts. Boosting, Random Forest, Random Tree shows significant results than bagging. In [3] a multi classifier was built using most promising classifiers for a given attack is validated for the KDD Cup dataset. The proposed multi-expert classifier showed better detection rate and FAR rate. As a result machine learning algorithm for intrusion detection process achieves significant result mainly for U2R and R2L attacks for the misuse detection concept. Here [4] the author presents boosting method to enable intrusion detection to detect possible intrusion of all systems without installing the software on client systems through web service using the IP address of the client system. The boosted decision tree is an alternative method to the existing techniques that have been used in IDS. In [5] compares four machine learning algorithms like J48, One R, Bayes Net and NB detect intrusion. Simulation results show that J48 decision tree is much suitable while compared with other three algorithms. It attains high TPR rate, low FTR rate and high accuracy results. Decision tree based SVM [6] solves multi-class problems. It also decreases the training and testing time with better efficiency results. Binary trees divide the dataset into two subsets from root to leaf until every subset has only one class. This construction plays a vital role in providing better classification efficiency. In [7] the author introduces an algorithm to adjust the weights of dataset based on probability and to split the data set into subset until all of its belong to a same class. Whereas the decision tree algorithm the weight of every example is set to equal value which differs from general characteristics. Simulation [8] have been done using decision tree algorithms like J48, AD, FT and LAD. Its corresponding TP rate, FP rate, precision , recall and F- measure are also evaluated to prove the efficiency of classification.

3 Performance Evaluation

Performance of tree based classifiers is evaluated using KDD Cup 99 dataset based upon following criteria,

3.1 10 Fold - Cross Validation

Cross-validation is also known as rotation estimation. It is a model to access the results of a statistical analysis to generalize an independent data set. It is mainly used in goal setting based on prediction. 10- Fold cross-validation is normally used in K-fold crossvalidation through which the folds are selected to estimate the mean response value is roughly equal in all the folds.

3.2 Comparison Criteria

The performance of the models is estimated using Accuracy, Sensitivity and Specificity. The accuracy, sensitivity and specificity were calculated by TP (True Positive) measure, FP (False Positive) measure, FN (False Negative) measure and TN (True Negative) measure [1].

Accuracy is the possibility that the algorithms can correctly predict positive and negative instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Sensitivity is the probability that the algorithms can correctly predict positive instances.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

Specificity is the probability that the algorithms can correctly predict negative instances.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

Mean absolute error is the average difference between the predicted and actual value in all test cases; it is the average prediction error. The formula for calculating MAE is given in equation shown below:

$$MAE = (|a1 - c1| + |a2 - c2| + \ldots + |an - cn|) / n \quad (4)$$

Assuming that the actual output as ‘a’ and probable output as ‘c’.

Root Mean Squared Error is frequently used values to predict a model or estimation technique through which the values are observed from the being modeled or estimated. The square root of the mean absolute error gives an RMSE which as follows,

$$RMSE = \sqrt{(a_1 - c_1)2 + (a_2 - c_2)2 + \cdots + (a_n - c_n)} / n \quad (5)$$

The mean squared error is used for numeric prediction. This value is computed by analyzing the average of the squared differences between computed value and its corresponding correct value. The accuracy of mean absolute and root mean squared error has been calculated for each machine learning algorithm.

4 Model Evaluation

Different Tree based Classifiers are used in this work in order to evaluate the effectiveness of these classifiers in a classification problem. The Classifiers applied are as follows:

4.1 Decision Stump

It is a machine learning model [9] composed of simple one level binary decision tree with an additional branch for missing values. It extends a third branch from the stump and prediction is done based on the value of a single input feature called as 1-rue. For nominal features, a stump includes a leaf for every possible feature value or two leaves one corresponds to chosen category and the other leaf to all the other categories. In case of binary features these two phenomenon are same. A missing value comes under another category. For continuous features, normally threshold value will be examined and stump contains two leaves for both minimum and maximum values of threshold. It also used as components with bagging and boosting capabilities. Decision stump is the kind of weak learner because it cannot give the best classification for the samples but a rather simple and fast classifier with accuracy at least just greater than 50% approximately.

4.2 Best First Tree

The best node split [10] leads to a maximum reduction of impurity among all nodes available for splitting. It constructs binary trees and attempts to maximize within node homogeneity. The extension of a node does not represent a homogeneous subset of cases seems to be an indication of impurity in which all cases have the same value for the dependent variable is a homogeneous node that requires no further splitting because it is pure. The impurity measures for nominal dependent variables are entropy based definitions of information gain and Gini index.

4.3 ID3

Iterative Dichotomiser 3 (ID3) method has been described using the information gain criterion is known as ID3 [11]. The use of gain ratio was one of many improvements that were made to ID3. Quinlal described it as robust under a wide variety of circumstances. ID3 stops if all attributes of training set classify successful and operates recursively on n, where n is the number of possible values of an attribute of the portioned subets to get their best attribute. Only one attribute at a time is tested for making a decision.

4.4 J48

It uses divide and conquer [5] algorithm to split a root node into possible subsets. Tree structure C can be formed using the following steps,

- The instances in class C belong to the same group or presence of fewer instances are labeled as a most frequent class in C.
- If not, choose a single attribute with at least two or more possible outcomes to form the root node of the tree which is followed by C1, C2, C3 and so on.
- Information gain is obtained to make the decision with highest normalization factor.

4.5 LAD Tree

Logical Analysis of Data builds [12] a classifier for binary target variable based on learning a logical expression so as to differentiate between positive and negative samples available in the dataset. It has been done through assumption where binary points cover only positive patterns and omits negative pattern tends to be positive. Similarly binary points that cover only negative pattern and excludes positive pattern is said to be negative. The construction of the LAD model for a given dataset handles the generation of a huge set of patterns and selection of a subset of them through prescribed assumption for every pattern in the model.

4.6 Random Tree

It chooses a test based on a given number of random features at each node performing no pruning. It is also specified as a tree or reassembling the structure that is formed by a stochastic process [13].

4.7 Reduced Error Pruning Tree

It builds a decision or a regression tree using information gain or variance reduction and prunes using a reduced error pruning method. It is a fast regression tree learner that uses information variance reduction in the data set which is spliced into a training set and a prune set [14]. Top-down induction of decision trees suffer from the inadequate functioning of the pruning phase.

4.8 Random Forest

It is constructed by bagging ensembles of random trees. Grow many classification trees using a probabilistic scheme. A random forest [15] of trees classifies a new object using an input vector by placing the input vector down each of the trees in the forest followed by classification based on total number votes over all the trees in the forest. Breiman's random forest technique blends elements of random subspaces and bagging in a way that is specific to using decision trees as the base classifier. Random

forests are prone to over fitting for some data sets. This is even more pronounced in noisy classification/regression tasks. Random forests do not handle large numbers of irrelevant features.

4.9 Simple Cart

It is a data exploration and prediction algorithm [16]. Classification and Regression trees are a classification method which is to construct decision trees uses historical data. CART may have unstable decision trees and splits only by one variable.

5 Result and Discussion

KDD CUP 99 Dataset [17] is developed based on DARPA 98 dataset in an MIT Lincoln Laboratory. Protocol such as TCP, UDP and ICMP has been used in this dataset to evaluate the anomaly detection methods. The dataset contains 24 different training attacks and 14 types in the test data. Here 7500 records are selected for the study out of 3, 11,029 Corrected dataset. Attacks such as Probe, U2R, and R2L are found to be less. 80% of data belongs to DoS attack respectively.

This work is performed using Machine learning tool to predict the effectiveness of all tree based Classifiers. The performance of the various algorithms is measured using classification accuracy, Sensitivity, Specificity, RMSE and MAE values. Table1 Comparison among tree based classifiers in terms of MAE, RMSE Accuracy, Sensitivity, Specificity and Time complexity. Figure1 (A) specifies the corresponding chart for the result obtained in table 1.

Table 1. Comparison Based on MAE, RMSE, Accuracy, Sensitivity, Specificity and Time

Algorithm	Accuracy (%)	MAE	RMSE	Sensitivity	Specificity	Time (Sec.)
Decision Stump	94.21	0.019	0.106	88.19	91.79	20.12
BF Tree	65.35	0.120	0.24	63.50	78.47	0.03
ID3	88.90	0.011	0.091	78.35	84.90	0.39
J48	93.21	0.020	0.110	81.25	87.74	0.19
LAD	95.45	0.016	0.101	92.18	93.54	10.97
Random Tree	94.70	0.016	0.01	90.20	92.10	0.08
REP Tree	93.14	0.022	0.115	87.08	94.01	0.58
Random Forest	**96.50**	**0.018**	**0.091**	**91.80**	**94.42**	**0.75**
Simple Cart	94.30	0.016	0.105	89.90	92.54	21.10

Figure 1 (B), illustrates the build time of all tree based classifiers. Simple CART Classifier consumes more time to build the model. Random tree, the probabilistic classifier tends to learn more rapidly for the given dataset. Since Simple Cart and Decision Stump take more time but also shows the accuracy rank two. Due to the high time variance of these two classifiers they are excluded from the graph (B). Random Forest even takes less time and better accuracy rate compared to all tree based classifiers.

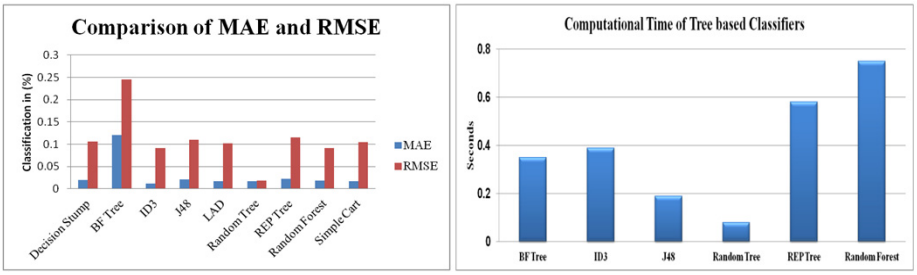


Fig. 1. (A): Graphical Representation of MAE & RMSE Values and (B): Time Complexity Measures of Tree Based Classifiers

Figure 2 (A) and (B) represents the accuracy, sensitivity and specificity values for all nine tree based classifiers. Based on values evaluated, accuracy of Decision Stump is 94.21%, the accuracy of BF Tree is 65.35%, the accuracy of ID3 is 88.90%, the accuracy of J48 is 93.21%, the accuracy of Least Error Prune is 95.45%, the accuracy of Random Tree is 94.70%, the accuracy of REP Tree is 93.14%, the accuracy of Random Forest is 96.50% and the accuracy of Simple Cart is 94.30%. Finally, Random Forest Classifier took highest accuracy percentage compared to 9 tree based classifiers.

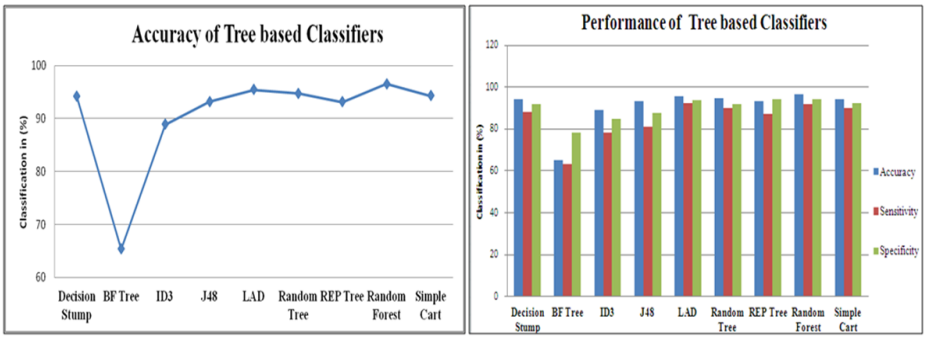


Fig. 2. (A) and (B): Comparison Based on Correctly Classified Instances, Specificity & Sensitivity

6 Conclusion and Future Enhancement

Due to the rapid growth of the network, new attack tends to happen. Intelligent IDS not only detect new kind of attacks but also has a low impact ratio. Machine learning algorithms can improve the efficiency of IDS. In this study, tree based classifiers are experimented to estimate the classification accuracy of that classifier in a classification problem. The experiment was done using an open source Machine Learning Tool. The performances of the classifiers were measured using 10-fold cross validation model and the results are compared using the KDD Cup Data set. Among nine classifiers (Decision Stump, BF Tree, ID3, J48, LAD, Random Tree, REP Tree, Random Forest and Simple Cart Classifiers), Random Forest Classifier performs well in the classification problem.

Random Tree classifier and Simple Cart Classifier are coming in the next category to classify the data. In future, ensembling of various data mining techniques has been deployed to develop a more efficient distributed Intrusion Detection System.

References

[1] Han, J., Kamber: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufman Publishers, Elsevier Inc. (2006)

[2] Banfield, R.E., Bowyer, K.W., Philip Kegelmeyer, W.: A Comparison of Decision Tree Ensemble Creation Techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–18 (2006)

[3] Sabhnani, M., Serpen, G.: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, pp. 1–7

[4] Renu Deepti, S., Loshma, G.: A Novel Data Mining Based Approach for Remote Intrusion Detection. International Journal of Computer Trends and Technology 3(3), 430–435 (2012)

[5] Kumar, Y., Upendra: An efficient Intrusion Detection Based on Detecision Tree Classifier Using Feature Reduction. International Journal of Scientific and Research Publications 2(1), 1–6 (2012)

[6] Mulay, S.A., Devale, P.R., Garje, G.V.: Intrusion Detection System using Support Vector Machine and Decision Tree. International Journal of Computer Applications 3(3), 40–43 (2010)

[7] Gaikwad, V.S., Kulkarni, P.J.: One Versus All Classification in Network Intrusion detection using decision tree. International Journal of Scientific and Research Publications 2(3), 1–5 (2012)

[8] Sharma, T.C., Jain, M.: WEKA Approach for Comparative Study of Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering 2(4), 1925–1931 (2013)

[9] Available on Wikipedia, `http://en.wikipedia.org/wiki/Decision_Stump` (last accessed on August 12)

[10] Kumar, N., Obi Reddy, G.P., Chatterji, S.: Evaluation of Best First Decision Tree on Categorical Soil Survey Data for Land Capability Classification. International Journal of Computer Applications 72(4), 5–8 (2013)

[11] Quinlan, J.R.: Induction of Decision Trees. Machine Learning (1), 81–106 (1986)

[12] Folorunsho, O.: Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database. International Journal of Advanced Research in Computer Science and Software Engineering 3(3), 11–15 (2013)

[13] `http://weka.sourceforge.net/doc/weka/classifiers/trees/RandomTree.html` (last accessed on August 12)

[14] Singh, S., Gupta, D.L., Malviya, A.K.: Performance Analysis of Classification Tree Learning Algorithms. International Journal of Computer Applications 55(6), 39–44 (2012)

[15] Breiman, L.: Random Forest. Machine Learning 45(1), 5–32 (2001)

[16] Sharma, A.K., Sahnip, S.: A Comparative Study of Classification Algorithms for Spam Email Data Analysis. International Journal on Computer Science and Engineering 3(5), 1890–1895 (2011)

[17] KDD Cup 99 intrusion Detection Data set, `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`

Heart Disease Classification Using PCA and Feed Forward Neural Networks

T. Santhanam[1] and E.P. Ephzibah[2]

[1] Department of Computer Science D.G. Vaishnav College, Arumbakkam, Chennai
[2] School of Information Technology and Engineering, VIT University, Vellore
santhanam_dgvc@yahoo.com,
ep.ephzibah@vit.ac.in

Abstract. The primary objective of this work is to discover a meaningful information in heart disease dataset for better diagnosis. This work is done using the data set available in UCI Machine learning repository. The work focuses on selecting the important features in the dataset using Principal Component Analysis and regression techniques. Using regression, the exponentiated estimate of the coefficient exp(B) of the feature is considered for feature selection. The exp(B) is the odds ratio of the independent variables. The work is done taking into consideration the components extracted using Principal Components Analysis technique and applying various operations on these components to produce methods like PCA1, PCA2, PCA3 and PCA4. It is observed that for one of the proposed methods PCA1, the prediction accuracy is 92.0% using regression and 95.2% using feed forward neural network classifier which is better than other methods. It is also observed that the accuracy of exp(B) is closer to PCA1 method, hence concluding that the exp(B) can also be considered for feature selection.

Keywords: Disease diagnosis, Principal Component Analysis, Feed Forward Neural Networks.

1 Introduction

Coronary heart disease is one of the leading causes of death in many of the developed, developing and under developed countries of the world. Heart failure is a condition in which the heart is unable to pump out enough oxygen-rich blood [1]. As the condition of the heart grows more serious in later stages it is better if the disease is diagnoised in the early stage.The disease can be diagnoised with the help of a number of tests. There are various factors that help the doctors to take desicions. They depend on the test results of the patients. The dataset under study is such a dataset where the test results of the patients are available. The class label that indicates whether the patient has heart disease or not, is also available. The proposed work provides a potentially useful decision making aid to the physicians for diagnosing the heart disease in patients. This decision making tasks can be done with a fewer number of tests. To retriev more

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 90–99, 2013.
© Springer International Publishing Switzerland 2013

meaningful information it is necessary to analyze and understand the data. The focus of this work is to findout the most important set of features for heart disease diagnosis system using Principal Component Analysis(PCA) and the exponentiated estimate of the coefficient exp(B) obtained using regression model.

According to the nobel prize winner Herbert A.Simon in 1978, "Learning is any process by which the performance is improved by experience". Machine learning is a methodology that trains the machine for proper decision making and to exhibit good performance. There are basically three different types of machine learning. They are supervised ,unsupervised and reinforcement learning. When machines are trainined they learn the pattern and try to take decisions based on their training and these systems are called as artificial intelligent systems. These systems when trained with and without the class labels are called as supervised learning and unsupervised learning respectively. There are many machine learning algortihms available for prediction and classification. Linear regression, Decision tree, Naive bayes, Neural networks, K Nearest neighbour are some of the machine learning algorithms[2]. These algorithms help to discover the hidden pattern available in the data and also to predict the output for the new incoming data. In unsupervised learning the clustering algorithms help in understanding the pattern available in the data and perform classification and prediction. In reinforcement learning there are agents that stochastically learn from the environment. The proposed work uses the Feed Forward Neural Network (FFNN) algorithm. It is a supervised learning method for classification where the connections in between the different layers do not form a cycle. Artificial neural networks provide good solutions for medical problems[3].

The performance of the system is reduced with increase in dimension of the data. Feature selection chooses a subset of features using an objective function. The selected or extracted features are used for classification using two approaches like filter approach and wrapper approach[4]. In filter approach the features are selected irrespective of the classifier used whereas in wrapper approach feature selection is done based on the classifier.

In section 2, a literature survey consisting of research works done related to disease diagnosis using various techniques is given. Section 3 provides the details about the features present in the heart disease dataset. Section 4 explains about factor analysis, principal component analysis, exponentiated estimate of the coefficient exp(B) obtained in regression. The section 5 provides the performance evaluation of the various methods used in this work. The section 6 concludes.

2 Literature Survey

Data mining techniques have been used for many decades. Still there are new techniques and tools emerging. Especially on disease diagnosis there are many number of research works been done. There are a few algorithms that are most commonly used for disease diagnosis like neural network algorithms [5, 6] regression techniques [7,8], decision tree algorithms [9] etc,. In [5] Orhan Er et al., have proved in their paper that probabilistic neural network is the best

classifier for Mesothelioma's disease with a prediction accuracy of 96.30% via 3 fold cross validation. The authors in [6] have proved that multi layer neural networks with two hidden layers work better with an average accuracy of 91.60 % for the diagnosis of chest diseases. Liu X et al., in [8] have used regression and Locally Linear Embedding (LLE) techniques for the classification of Alzhemier's disease and have got the sensitivity of about 80%. Feature selection is a part of their approach in [9], for the heart and hepatitis datasets. Using C4.5 decision tree algorithm the authors have selected features and evaluated the accuracies as 92.59% and 81.82% for heart and hepatitis disease diagnosis respectively. Searching for an optimal feature subset is a NP-complete problem [10]. The SAGA method in [10] is a hybrid algorithm that combines simulated annealing, genetic algorithms, generalized regression neural networks and greedy algorithms for feature selection in large dimensionality. The results show that the performance of SAGA in terms of accuracy is 93.0%.

3 Heart Disease Dataset

The dataset for heart disease is taken from the UCI Machine Learning Repository [11]. The data set consists of 303 samples from which 6 samples are removed due to the presence of missing values. One of the ways of data cleaning is ignoring the missing entries as in [14]. The proposed work has eliminated the samples that contain missing entries as in other famous research paper[9]. Therefore a total of 297 samples are considered for further processing. There are 13 input and one output attribute. The output attribute contains 5 classes ranging from 0 to 4. The value 0 represents the samples without heart disease and values from 1 to 4 represents the presence of the disease with the level of its severity from lowest to highest respectively. In this work the class label is assigned a binary value that can be either 0 or 1 [15,16]. The features in the dataset are Age, Sex, Chest pain type, Resting blood pressure, Serum cholesterol, Fasting blood sugar, Rest ECG, Thalach- Maximum heart rate achieved, Exercise induced angina, Old Peak, Slope, Number of colored vessels, Thal and the class label.

4 Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors [13]. This helps us to understand the structure of data, for further processing. The first step in factor analysis is the identification of the domain and population of interest. In the proposed study the domain is medical field and the population is the data set associated with it. Population is a large volume of data which is very difficult to handle. Therefore it is necessary to identify the required set of attributes and samples that are to be measured from the population. These are called surface attributes. There are also a set of internal variables called factors that makes them different from other attributes.

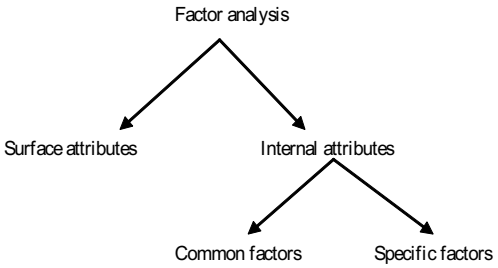


Fig. 1. Attributes and Factors in Common factor theory

They are common or specific based on whether they affect more than one of the surfaces attributes respectively.

Always attributes tend to correlate with each other. It is because of the effect of common factors. The specific factors do not contribute anything towards correlation among surface attributes. Any system cannot be perfect and always there is more possibilities of errors accounted. They are a part of the system and are called as errors of measurement. There are two types of factor analysis based on how they represent the correlation or interrelationship among the attributes. Exploratory Factor Analysis (EFA) helps us to understand the complex relationship among the attributes. Confirmatory Factor Analysis (CFA) tests the hypothesis[13]. The proposed work is done using EFA method and identifies the relationship between the features of the heart disease dataset.

4.1 Principal Component Analysis (PCA)

A meaningful relationship in the data is called a pattern. PCA is one of the factor analysis techniques that identifies a meaningful interrelationship among the attributes in the data. It provides a way of obtaining data with a reduced dimension without any loss. Standard deviation of a dataset can be evaluated as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{1}$$

s is a measure that contains the average distance between the mean of the dataset to a point in the set. Variance is identical to the standard deviation. The formula given below can be used to calculate the variance in the data.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{2}$$

It is possible to estimate these measures for a single dimension independently, where as to find the interrelationship among the attributes, covariance is used.

Covariance computes the similarity between any two attributes. The following formula is used for this purpose.

$$cov\,(X,Y) = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1} \tag{3}$$

PCA identifies the components using Eigenvectors of the covariance or correlation matrix. Let A be a complex square matrix. Then if λ is a complex number and X a non-zero complex column vector satisfying AX= λ X, X is an eigenvector of A, where λ is called an eigenvalue of A. X is an eigenvector corresponding to the eigenvalue λ. PCA helps us to find out factors that explain the pattern of correlations within a set of observed variables. In the heart disease dataset there are 13 input attributes. The PCA applied on the data reduces the number of features from 13 to 7±3 and also groups them under 4 components as in table(1). The scree plot is used to plot the components according to their Eigen values.

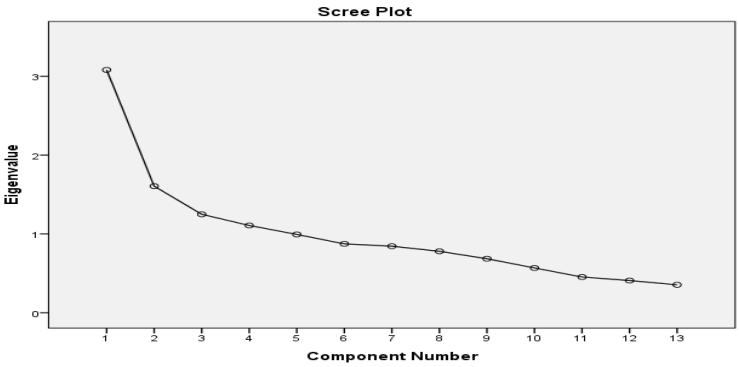


Fig. 2. Scree Plot

The components are extracted from the steep slope rather than shallow slope as the components in the shallow slope contribute little to the solution. There is a drop between the second and third component in the above scree plot. Hence selecting the first two components is a better choice. Therefore only two components are considered for further analysis. The table 1 given below gives the detail about the number of extracted components based on their Eigen values. Thus, indicating that there are four major components in the data set. The first section (sec1) of the table gives the total, % of variance and the cumulative % of the Eigen values for all the features in the dataset. The % of variance column gives the ratio of variance to the total by all variables. The cumulative % column gives the % of variance accounted for by the first n-1 components for the nth component, where n represents the feature number. The cumulative % value is calculated by the sum of the % of variance for the n-1 and n-2 components. The next section (sec2) shows the details about the extracted sums of the squared

Table 1. Total Variance

Comp	Initial Eigen values Sec(1)			Sums of Squared Loadings Sec(2)		
Comp	Total	%ofVariance	Cumulative %	Total	%ofVariance	Cumulative%
1	3.080	23.695	23.695	3.080	23.695	23.695
2	1.605	12.349	36.045	1.605	12.349	36.045
3	1.248	9.604	45.648	1.248	9.604	45.648
4	1.107	8.516	54.165	1.107	8.516	54.165
5	.993	7.638	61.803			
6	.874	6.720	68.523			
7	.844	6.494	75.017			
8	.779	5.994	81.011			
9	.685	5.269	86.280			
10	.568	4.368	90.648			
11	.453	3.486	94.135			
12	.408	3.140	97.275			
13	.354	2.725	100.00			

loadings. This section indicates that the extracted components have approximately 54% of variability in the original set of values and there is around 46 % loss of information.

The component table (table 2) is obtained by extracting the first two components. The first component is highly correlated with the attributes like ST depression scale, slope of the peak, thal, exercise induced angina, number of major vessels colored, chest pain type. The second component is highly correlated with the attributes like serum cholesterol, resting blood pressure. The attribute max heart rate achieved and sex have a negative correlation with component 1 and 2 respectively and so they are rejected. The attribute age has the correlation in both the components hence it is rejected. Resting ECG results and fasting blood sugar are weakly correlated with the component 2. The features are said to have weak correlations with the components if their correlation value is lesser than 0.4[17]. Hence they are also eliminated from the dataset. As feature selection is done using PCA, there are some methods designed and proposed using PCA. Those methods are named as PCA1, PCA2, PCA3 and PCA4. PCA1 is a method that takes into account the component values of the features and their sum. The feature that has the highest sum gets rank1 and other features are arranged based on their values in descending order. The methods PCA2 and PCA3 are similar to PCA1 but instead of summing the values the component values were subtracted and multiplied respectively. For PCA4 method the maximum value from the set of components is obtained and names as M. For all the features in the dataset the difference between M and their corresponding component value is calculated. Again the features are arranged in descending order and ranked. Features are selected based on their ranks. The selected features using PCA method were used for classification and prediction using regression and FFNN classifiers.

Table 2. Component Matrix

Features	Comp 1	Comp 2
ST depression scale	.697	
Max Heartrate achieved	-.689	
Slope of the peak	.618	
thal	.608	-.334
Exercise induced angina	.585	
Num of major vessels colored	.538	
Chest pain type	.502	
Sex		-.547
Serum cholestrol		.543
Age .	502	.530
Resting blood pressure		.496
Resting ecg results		.338
Fasting blood sugar		.304

4.2 Regression Analysis

Regression is a statistical process that calculates the relationship among the features in a dataset. Regression analysis helps to identify the change in the dependent variable (i.e., the class label) when there is a slight modification in any one of the independent features. Regression can be done using different methods based on the characteristics of the independent features. If all the independent features pass the test of normality then a linear regression can be done on the data. But, if any one of the independent features fail to pass the test of normality then the logistic regression can be done. If there are two categories in the dependent variable the classification is called as binary regression, whereas for multiple numbers of classes the multinomial regression is the appropriate method. Table 3 is obtained by performing binary logistic regression on the heart disease dataset as there are binary values in the class label and also all the features have not passed the test of normality.

The details like the regression coefficient (B), standard error (S.E), Wald statistic, significance level (sig), exponentiated estimate of the coefficient exp(B) and the confidence interval for exp(B) are available in table 3. The focus is on the exp(B) value that represents the change in the attributes for one unit change in the predictor. With a threshold set to be greater than 1.2 units, the attributes are selected. The exp(B) column helps us to identify the predictive capability of each and every attribute. It is found that the predictive capability values are highest and lowest in sex and fasting blood sugar respectively. The exp(B) value interprets the weight of the feature in prediction. The following equation gives the logistic regression model:

$$logit(p) = \log \frac{p}{1-p} \tag{4}$$

Table 3. Identifying the features using EXP(B)

Features	B	S.E	Wald	Sig	Exp(B)	95%C.I for Exp(B)(upper)
age	-0.014	.024	.349	.555	.986	1.033
sex	1.312	.488	7.215	.007	3.714	9.674
cp	.576	.191	9.073	.003	1.779	2.587
fbs	.024	.011	5.021	.025	1.024	1.046
chol	.005	.004	1.752	.186	1.005	1.012
fbs	-1.022	.555	3.386	.066	.360	1.069
restecg	.245	.185	1.756	.185	1.278	1.836
thalach	-.021	.010	4.085	.043	.980	.999
exang	.926	.413	5.020	.025	2.525	5.676
oldpeak	.247	.212	1.364	.243	1.281	1.940
slope	.570	.363	2.465	.116	1.768	3.603
ca	1.268	.265	22.819	.000	3.553	5.977
thal	.344	.100	11.744	.001	1.410	1.717

$$\log \frac{\hat{p}_i}{1-\hat{p}_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{5}$$

By exponentiating both sides we get odds ratio exp(B).

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \ldots e^{\beta_p x_{ip}} \tag{6}$$

Here, β is the regression coefficient and the exponential function of the regression coefficient e^{β} is the odds ratio.

Table 4. Features selected using the proposed methods

Features	PCA	PCA1	PCA2	PCA3	PC4	exp(B)
Age			1		1	
Sex		1		1		1
Chest pain type	1	1	1	1	1	
Trespbps	1		1			
Cholesterol	1		1			
Fasting blood sugar			1			
Rest ECG			1			1
Thalach					1	
Exang	1	1	1	1	1	1
Oldpeak	1	1	1	1	1	1
Slope	1	1	1	1	1	1
Ca	1	1	1		1	1
Thal	1	1	1	1	1	1

Table 5. Results based on the proposed work

SNo	Methods	TP in reg	Features	FFNN
1	All features	90.0	13	90.54
2	PCA	88.5	8±1	87.6
3	PCA1	92.0	7±1	95.2
4	PCA2	83.5	10±1	82.5
5	PCA3	85.5	7±1	86.5
6	PCA4	88.0	7±1	88.2
7	EXP(B)	87.5	8±1	85.2

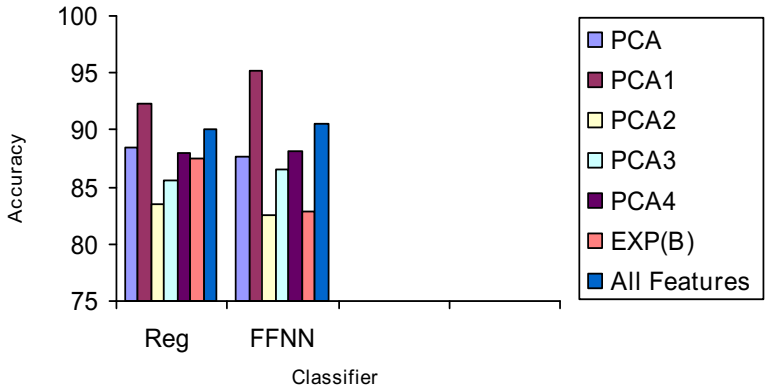


Fig. 3. Performance Evaluation

5 Performance Evaluation

Table 4 depicts the features that are selected based on the methods PCA, PCA1, PCA2, PCA3, PCA4 and exp(B) methods. The table 5 shows the classification accuracy obtained using PCA methods and exp(B) with reduced set of features using MATLAB version 7.10. It is identified that the exponentiated estimate of the coefficient exp(B)is also used to select features and its accuracy is 87.5% and 85.2 % which is 4.5% and 5.0%lesser than the accuracy of the PCA using regression and FFNN classifiers respectively. In table 5 reg represents regression and TP for Ture Positive. The Accuracy of the results is in percentage. In figure 3 the performance of the methods are represented graphically.

6 Conclusion

Data analysis plays an important role in classification and prediction models. In the proposed work feature selection is done using the methods like PCA and exponentiated estimate of the coefficient exp(B), in regression. It is evaluated that PCA and exp(B) give almost equal prediction accuracy with a difference of 2.0%. Hence it can be stated that exp(B) is also a method suitable for feature

selection. Among the PCA methods the PCA1 gives the maximum accuracy of 92.0% and 95.2% using regression and FFNN classifiers respectively. The scope of the future work will explore the possibility of using other NN models for disease classification.

References

1. A.D.A.M. Medical Encyclopedia, Heart failure overview. PubMed Health (2013)
2. Hassanien, A.E., Al-Shammari, E.T., Ghali, N.I.: Computational intelligence techniques in bioinformatics. Computational Biology and Chemistry 47, 37–47 (2013)
3. Ghumbre, S.U., Ghatol, A.A.: An intelligent system for hepatitis b disease diagnosis. International Journal of Computers and Applications 32(4), 455–460 (2010)
4. Kung, S.Y., Luo, Y., Mak, M.-W.: Feature Selection for Genomic Signal Processing: Unsupervised, Supervised, and Self-Supervised Scenarios. J. Sign. Process. Syst., 3–20 (2010)
5. Er., O., Temurtas, F., Cetin Tanrikulu, A.: An approach on probabilistic neural network for diagnosis of mesothelioma's disease. Computers and Electrical Engineering, 75–81 (2012)
6. Er., O., Yumusak, N., Temurtas, F.: Chest diseases diagnosis using artificial neural networks. Expert Systems with Applications, 7648–7655 (2010)
7. Shao, Y.E., Hou, C.-D., Chan, Y.-C.: The hybrid logistics regression-artificial neural network and multivariate adaptive regression splines-artificial neural network modeling schemes for heart disease classification. Advanced Science Letters 19(11), 3405–3408 (2013)
8. Liv, X., Tosun, D., Weiner, M.W., Schuff, N.: Locally linear embedding for MRI based Alzhemier's disease classification. NeuroImage 83, 148–157 (2013)
9. Polat, K., Gunes, S.: A hybrid approach to medical decision support systems:Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer Methods and Programs in Biomedicine, 164–174 (2007)
10. Gheyas, I.A., Smith, L.S.: Feature subset selection in large dimensionality domains. Pattern Recognition 43, 5–13 (2010)
11. Detrano, R.: V.A. Medical Center Long Each and Cleveland Clinic Foundation, `ww.archive.ics.uci.edu/ml/datasets`
12. Tucker, L.R., MacCallum, R.C.: Exploratory factor analysis (1997)
13. Han, J., Kamber, M.: Data Mining Concepts and Techniques, p.109 (2001)
14. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques, pp. 108–115. IEEE (2008)
15. Polat, K., Gunes, S.: A new feature selection method on classification of medical datasets: Kernel F-Score feature selection. Expert Systems with Applications, 10367–10373 (2009)
16. Lee, K., Ahn, H., Moon, H., Kodell, R.L., Chen, J.J.: Multinomial logistic regression ensembles. PubMed (2013)
17. Abawajy, J.H., Kelarev, A.V., Chowdhury, M.: Multistage approach for clustering and classification of ECG data. Computer Methods and Programs in Biomedicine 1–11 (2013)

Non-disjoint Cluster Analysis with Non-uniform Density

Chiheb-Eddine Ben N'Cir and Nadia Essoussi

LARODEC, ISG of Tunis, University of Tunis, Tunisia
{Chiheb.benncir,Nadia.essoussi}@isg.rnu.tn

Abstract. Non-disjoint clustering, also referred to as overlapping clustering, is a challenging issue in clustering which allows an observation to belong to more than one cluster. Several overlapping methods were proposed to solve this issue. Although the effectiveness of these methods to build non-disjoint partitioning, they usually fail when clusters have different densities. In order to detect overlapping clusters with uneven densities, we propose two clustering methods based on a new optimized criterion that incorporates the distance variation in a cluster to regularize the distance between a data point and the cluster representative. Experiments performed on simulated data and real world benchmarks show that proposed methods have better performance, compared to existing ones, when clusters have different densities.

Keywords: Overlapping Clustering, Clusters with Different Densities, Overlapping k-means, Distance Variation.

1 Introduction

Given the explosion of data available from the web, from scientific research, from companies and social networks, developing effective tools for analyzing data efficiently has become increasingly important. In this way, clustering has become an important technique in data mining and pattern recognition to predict and to summarize data. The problem of clustering data has led to many methods and models which are widely used across many fields. Unfortunately, while clustering methods are a wonderful tool for many applications, they are actually quite limited. Clustering methods traditionally assume that each data point belongs to one and only one cluster leading to k exhaustive and disjoint clusters explaining the data. In many situations the data being modeled can have a much richer and more complex hidden representation than this disjoint one. For example, there may be overlapping regions where data points actually belong to multiple clusters. Solution to this problem contributes to solve many real problems that require to find overlapping regions in order to fit the data set structure. For example, in video classification, overlapping clustering is a necessary requirement where videos have potentially multiple genres [1]. In emotion detection, overlapping clustering methods need to detect different emotions for a specific piece of music [2], etc. A trend in the implementation of clustering methods called

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 100–111, 2013.
© Springer International Publishing Switzerland 2013

"overlapping clustering" has tried to solve these problems which have arisen naturally.

In fact, overlapping clustering is based on the assumption that each data point may belong to one or several clusters without any membership coefficient. Several overlapping clustering methods based on hierarchical, graph based, correlation and partitioning approaches are proposed in the literature. Our work distinguishes from this body of research as it develops within the k-means algorithm [3,4]. First proposed methods extend results of uncertain partitioning methods, such as results of fuzzy c-means[5], possibilistic c-means [6] and evidential c-means [7] to obtain overlapping clusters. These methods need a post-processing treatment to generate the final overlapping clusters by thresholding clusters memberships. More effective methods are based on the generalization of the optimized criterion of k-means to look for optimal non-disjoint partitioning. Example of these methods are OKM (Overlapping K-Means) [8], ALS (Alternating Least Square) [9,10], Parameterized R-OKM (Parameterized Restricted Overlapping K-means) [11]. As opposed to uncertain partitioning methods, these methods produce crisp overlapping clusters and do not need any post processing treatment.

Although the effectiveness of existing methods to build non disjoint groups, many suffer from some serious limitations. The currently existing overlapping methods consider only the Euclidean distance between each data point and its representative. Furthermore, the density of data points in a cluster could be distinctly different from other clusters in the data set. The used Euclidean metric evaluates only the distance between two individual data points and ignores the global distance variation for all data points in a cluster; thus fail to perform high quality clusters when data contain groups with different densities.

The goal of this paper is to design more efficient and perfective methods for overlapping clustering which address the limitation of producing relevant overlapping clusters when data contain groups with different densities. We propose two methods referred to as OKM-σ and Parameterized R-OKM-σ in which we incorporate the distance variation in a cluster into the optimized criterion to regularize the distance between a data point and its representative.

The remainder of this paper is organized as follows: Section 2 recalls the necessary background about existing overlapping methods based k-means algorithm. Then Section 3 presents the issue of identifying clusters with different densities. After that, Section 4 describes the proposed methods OKM-σ and Parameterized R-OKM-σ while Section 5 describes experiments performed in artificial and real data sets. Finally Section 6 presents the conclusion and the future works.

2 Overlapping Clustering

We describe in the following two existing overlapping methods based k-means algorithm which are respectively OKM and Parameterized R-OKM. These two methods are considered as a generalization of k-means for overlapping clustering.

2.1 OKM

OKM is an extension of k-means to detect optimal non disjoint clusters. This method proposes an extension of the objective function of k-means where overlaps are introduced and optimized iteratively. Given a set of N data points, OKM aims to find the optimal K partitions $\Pi = \{\pi_c\}_{c=1}^{K}$ such that the following objective function is optimized:

$$J(\Pi) = \sum_{i=1}^{N} ||x_i - (\overline{x_i})||^2 \tag{1}$$

where $\overline{x_i}$ is the average of representatives (denoted as "image") of clusters to which data point x_i belongs to:

$$\overline{x_i} = \sum_{k \in \Pi_i} \frac{c_k}{|\Pi_i|} \tag{2}$$

where c_k is the representative of cluster k and Π_i is the set of clusters to which data point x_i belongs to.

The minimization of the objective function is performed by iterating two independent steps: (1)computation of cluster representatives (C) and (2) multi assignment of data to one or several clusters (Π). The stopping rule of OKM algorithm is characterized by two criteria: the maximum number of iterations or the minimum improvement of the objective function between two iterations.

OKM has the issue that obtained overlaps between clusters are large. Known that overlapping clustering reconsiders the "well separated clusters" property, clusters with too large overlaps are not appropriate for most of the target applications. To solve this issue, a recent method, referred to as Parameterized R-OKM, proposes a new model which performs non disjoint clusters with control of overlaps.

2.2 Parameterized R-OKM

In order to produce clusters with acceptable overlaps' size, Parameterized R-OKM restricts the excessive assignments of a data point x_i to multiple clusters according to the cardinality of the set of assignments $|\Pi_i|$. Parameterized R-OKM is based on the minimization of the following objective criterion:

$$J(\Pi) = \sum_{i=1}^{N} |\Pi_i|^{\alpha}.||x_i - (\overline{x_i})||^2 \tag{3}$$

where $\alpha \geq 0$ is a parameter, fixed by the user, to control the size of the overlaps. As well as α becomes large, Parameterized R-OKM builds clusters with more reduced overlaps. When $\alpha = 0$, Parameterized R-OKM coincides exactly with OKM. For the minimization of the objective criterion, Parameterized R-OKM uses the same minimization steps used for OKM.

3 Clusters with Non Uniform Densities

In real life applications of overlapping clustering the density of data points in a cluster could be distinctly different from other clusters in the data set. Fig.1 presents the issue of groups with uneven densities and shows two classes with different densities where the red class have a low density compared to the blue class.

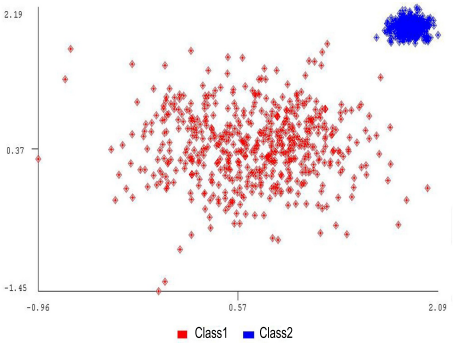


Fig. 1. Two classes with different densities: the "Red" class is characterized by low density and the "Blue" class is characterized by high density

To study patterns produced by existing overlapping methods when data have groups with different densities, we visualize the partitioning obtained by OKM and Parameterized R-OKM with 2 clusters in the data set described below. Fig.2 shows that OKM builds clusters with too large overlaps (the "Green" data points) and assigns some data points laying in the extremity surface between the two classes to the low density class (Red data points). For Parameterized R-OKM we show that overlaps are considerably reduced. However, data points in the extremity surface between the two obtained clusters are miss assigned.

In fact, OKM and Parameterized R-OKM consider only the Euclidean distance between each data point and each cluster representative. The used Euclidean metric evaluates only the distance between two individual data points and it ignores the global distance variation for all data points in a cluster. Therefore, OKM and Parameterized R-OKM fail to detect clusters with different densities leading to clusters with large overlaps. We show in the next section how can we design more efficient method to detect these types of clusters.

Recently, the problem of detecting groups with different densities is solved by using a new distance metric that incorporates the distance variation in a cluster to regularize the distance between a data point and the cluster's representative. This technique were used within Fuzzy c-means and referred to as Fuzzy c-means-σ (FCM-σ)[12]. This methods introduces the distance variation of each individual data group to regularize the distance between a data point and the cluster's centroid. This distance metric can be better applied to data with uneven

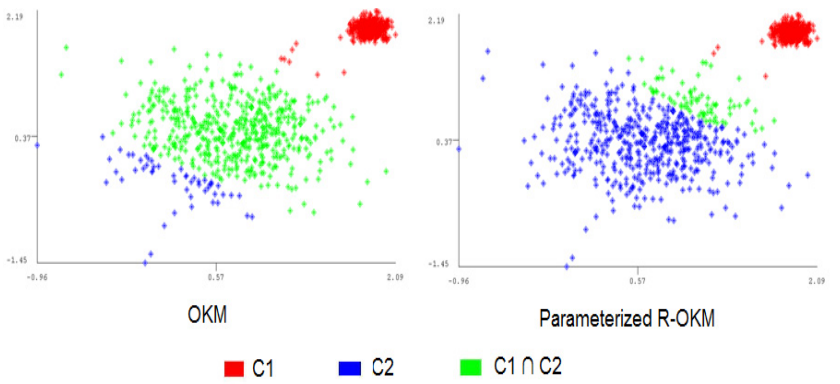


Fig. 2. Partitioning obtained by OKM and Parametrized R-OKM in an artificial data set having two classes with different densities

densities. The new distance metric between each data point x_i and each cluster's representative c_k is defined by:

$$\hat{d}^2_{ik} = \frac{||x_i - c_k||^2}{\sigma_k}, \tag{4}$$

where σ_k is the weighted mean distance in cluster k, and is given by:

$$\sigma_k = \left\{ \frac{\sum_{i=1}^{N} w_{ik}^p . ||x_i - c_k||^2}{\sum_{i=1}^{N} w_{ik}^p} \right\}^{1/2}, \tag{5}$$

where $w_{ik} \in [0, 1]$ indicates the membership degree that observation x_i belongs to cluster k and p is a parameter used to control fuzziness.

The evaluation of FCM-σ in data with equilibrated densities shows that this method gives the same results of the original FCM [12]. However, if data contain different groups with different densities, FCM-σ performs better than the original FCM.

4 Proposed Methods

In order to fit the data set structure, overlapping clustering methods should be able to detect non disjoint clusters with non uniform densities. Therefore, we propose two new methods, referred to as OKM-σ and Parameterized R-OKM-σ, which are able to produce more perfective patterns than existing methods when data have groups with uneven densities. The proposed methods introduce a regularization of clusters' density in their optimized criterion.

4.1 Overlapping k-means-σ (OKM-σ)

To take into account that clusters can have non uniform densities, OKM-σ introduces a regularization factor σ_i for each observation x_i. Given the set of N data points, the proposed method minimizes the following objective criterion:

$$\begin{aligned} J(\Pi) &= \sum_{i=1}^{N} \hat{d}^2(x_i, (\overline{x_i})) \\ &= \sum_{i=1}^{N} \frac{\parallel x_i - (\overline{x_i}) \parallel^2}{\sigma_i}, \end{aligned} \tag{6}$$

where σ_i is defined by the average of clusters' density σ_k to which x_i belongs to:

$$\sigma_i = \sum_{k \in \Pi_i} \frac{\sigma_k}{|\Pi_i|}. \tag{7}$$

To take into account that data can be assigned to more than cluster and to guarantee the decrease of the objective criterion, the cluster density σ_k is defined as the average distances between each data point x_i and its image $\overline{x_i}$ belonging to cluster k:

$$\sigma_k = \left\{ \frac{\sum_{i=1}^{N} P_{ik}.||x_i - (x_i)||^2}{\sum_{i=1}^{N} P_{ik}} \right\}^{1/2}, \tag{8}$$

where P_{ik} is a binary variable indicating membership of data point x_i to cluster k.

4.2 Parameterized R-OKM-σ

In the same way that OKM-σ, we introduce the distance based cluster density within the objective criterion of Parameterized R-OKM-σ. The new objective criterion is defined by:

$$\begin{aligned} J(\Pi) &= \sum_{i=1}^{N} |\Pi_i|^{\alpha} \hat{d}^2(x_i, (\overline{x_i})) \\ &= \sum_{i=1}^{N} |\Pi_i|^{\alpha} \frac{\parallel x_i - (\overline{x_i}) \parallel^2}{\sigma_i}, \end{aligned} \tag{9}$$

where σ_i is the regularization factor local to x_i as described in Eq.7 for OKM-σ. However, the cluster density σ_k is defined for Parameterized R-OKMσ by:

$$\sigma_k = \left\{ \frac{\sum_{i=1}^{N} P_{ik}.|\Pi_i|^{\alpha}.||x_i - (x_i)||^2}{\sum_{i=1}^{N} P_{ik}.|\Pi_i|^{\alpha}} \right\}^{1/2}. \tag{10}$$

4.3 Algorithmic Resolution

The minimization of the objective function of each proposed method (OKM-σ and Parameterized R-OKM-σ) is performed by iterating three steps: (1) computation of cluster representatives C, (2) multi assignment (Π) of observations to one or several clusters and (3) computation of weights (σ_k) for each cluster.

Algorithm 1. *Parameterized R-OKM-σ* $(X, t_{max}, \varepsilon, K) \rightarrow \Pi$

Require: X: a data set described over $\mathbb{R}^d$.
t_{max}: maximum number of iterations.
ε: minimal improvement in the objective function.
K: number of clusters.
Ensure: Π: assignment of observations over K clusters.
1: Initialize representatives of clusters C^0 randomly over X, initialize weights σ_k^0, initialize clusters memberships Π^0 using $ASSIGN{-}\sigma$ and compute the objective function $J(\Pi^0, C^0, \sigma^0)$ at iteration 0.
2: $t = t + 1$.
3: Update clusters representatives C^t.
4: Compute new assignments Π^t using $ASSIGN - \sigma(x_i, C^t, \Pi_i^{t-1})\ \forall i$.
5: Update weights σ^t using Eq.10).
6: Compute objective function $J(\Pi^t, C^t, \sigma^t)$.
7: **if** $(t < t_{max}$ and $J(\Pi^{t-1}, C^{t-1}, \sigma^{t-1}) - J(\Pi^t, C^t, \sigma^t) > \varepsilon)$ **then**
8: Go to step 2.
9: **else**
10: Return Π^t the final cluster memberships matrix.
11: **end if**

Known that OKM-σ is a specific case of Parameterized R-OKM-σ (when $\alpha = 0$), we present in Algorithm 1 the generic algorithm of Parameterized R-OKM-σ. This algorithm uses the function $ASSIGN{-}\sigma$ that defines the assignment strategy. This strategy consists, for each x_i, in sorting clusters from closest to farthest with respect to the new distance then assigning clusters in the order defined while assignment improves the local error $|\Pi_i|^\alpha \hat{d}^2(x_i, (\overline{x_i}))$ therefore, reducing the objective criterion after each assignment step. A pseudo code for the $ASSIGN{-}\sigma$ function is described in Algorithm 2.

For the step of prototype computation, using the lagrange multipliers method, by differentiating with respect to c_k and setting derivative to zero, optimal prototypes to made the objective criterion to be minimized are described by:

$$c_k = \frac{\sum_{x_i \in \pi_k} \frac{\sigma_i}{|\Pi_i|^{2-\alpha}} . \tilde{x_i}^k}{\sum_{x_i \in \pi_k} \frac{\sigma_i}{|\Pi_i|^{2-\alpha}}}, \tag{11}$$

where $\tilde{x_i}^k$ denotes representative c_k according to x_i such that $\parallel x_i - im_{\Pi,C}(x_i) \parallel^2 = 0$ and is computed by

$$\tilde{x_i}^k = |\Pi_i|.x_i - (|\Pi_i| - 1).im_{\Pi \setminus k,C}(x_i). \tag{12}$$

Algorithm 2. $ASSIGN - \sigma(x_i, \{c_1, ...c_K\}, \Pi_i^{old}) \rightarrow \Pi_i$

Require: x_i: Vector in $\mathbb{R}^d$.
$\{c_1, ...c_K\}$: K cluster representatives.
Π_i^{old}: Old assignment for observation x_i.
Ensure: Π_i: New assignment for x_i.
1: Initialize $\Pi_i = \{c^\star\}$ the nearest cluster where $c^\star = \arg min_{c_k} \parallel x_i - c_k \parallel^2$.
2: Looking for the next nearest cluster $c^\star$ which is not included in Π_i.
3: Compute $(\overline{x_i}')$ and σ_i' with assignments $\Pi_i' = \Pi_i \cup \{c^\star\}$.
4: **if** $\frac{|\Pi_i'|^\alpha . \|x_i - \overline{x_i}'\|^2}{\sigma_i'} \leq \frac{|\Pi_i|^\alpha . \|x_i - \overline{x_i}\|^2}{\sigma_i}$ **then**
5: $\Pi_i \leftarrow \Pi_i'$ and go to step 2.
6: **if** $\sigma_i . \parallel x_i - \overline{x_i} \parallel^2 \leq \sigma_i^{old} . \parallel x_i - \overline{x_i} \parallel^2$ **then**
7: Return Π_i.
8: **else**
9: Return Π_i^{old}.
10: **end if**
11: **end if**

5 Experiments

This section describes the details of the experiments that we performed in order to evaluate the ability of the new proposed methods to produce more relevant clusters when data contain groups with different densities. The evaluation is based on two fold assessment: firstly, as an "internal" point of view we give visual information about the patterns produced on artificial examples; then we perform a standard "external" evaluation by comparing clusterings obtained with the benchmark basis.

5.1 Internal Assessment

We simulated 2 artificial data sets, "Artificial data set 1" and "Artificial data set 2", containing two classes where each class contains 500 data points defined in two dimensional space as shown in Fig. 1. These two classes have different densities: the "Blue" class have high density than the "Red" class. For the second data set, we modified the radius of the "Red" class which becomes more large than the radius of this class in the first data set.

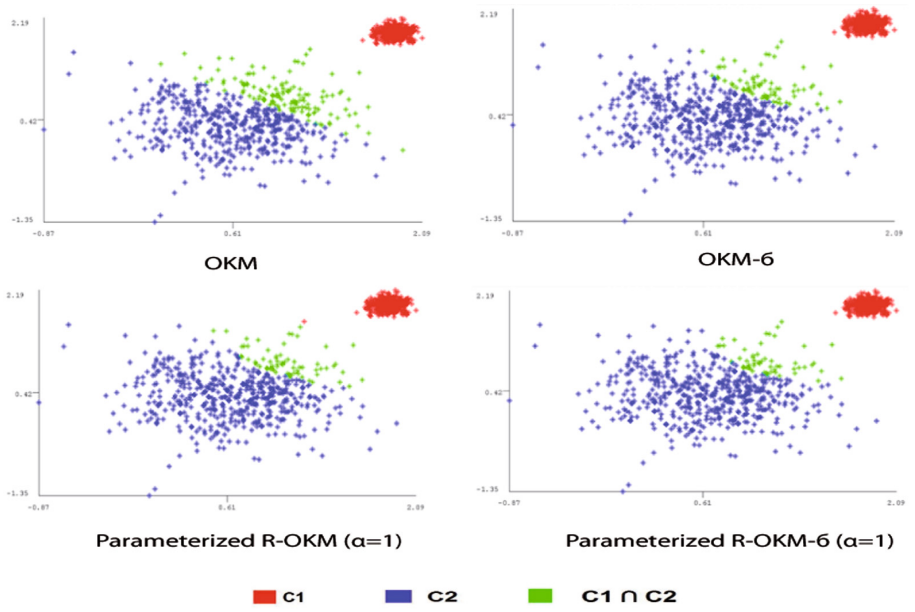


Fig. 3. Comparison of clusters obtained with OKM-σ and Parameterized R-OKM-σ versus clusters obtained with OKM and Parameterized R-OKM in "Artificial data set 1"

Figure 3 and Figure 4 show the resulting patterns using OKM-σ and Parameterized R-OKM-σ versus resulting patterns using existing methods. At first, we notice that all methods are able to produce non disjoint clusters. The "Green" data points represent data which are assigned to the intersection of cluster 1 and cluster 2. For OKM, we notice the large overlapping boundaries built by this method which explains the large surface of "Green" data points. This problem is partially solved using OKM-σ where overlaps between the two clusters are reduced in both data sets. The use of the distance variation in OKM-σ considerably reduces overlaps between the two clusters. For parameterized R-OKM, we notice that overlaps are reduced. However, the Parameterized R-OKM has the issue that some data points laying in the extremity surface between the two clusters are miss assigned: although that these data points was actually simulated within the cluster having the low density, these data points are assigned to the cluster having the high density (assigned to the "Red" cluster). The number of miss assigned data points increases as the difference of densities between clusters increases which explains the increase of miss assigned data points in the second data set as shown in Fig. 4 (7 miss assigned data points in "Artificial data set 2" versus 1 miss assigned data point in "Artificial data set 1"). This problem is solved when we used Parameterized R-OKM-σ. The produced patterns are more relevant and fit better the actual structure of the simulated data.

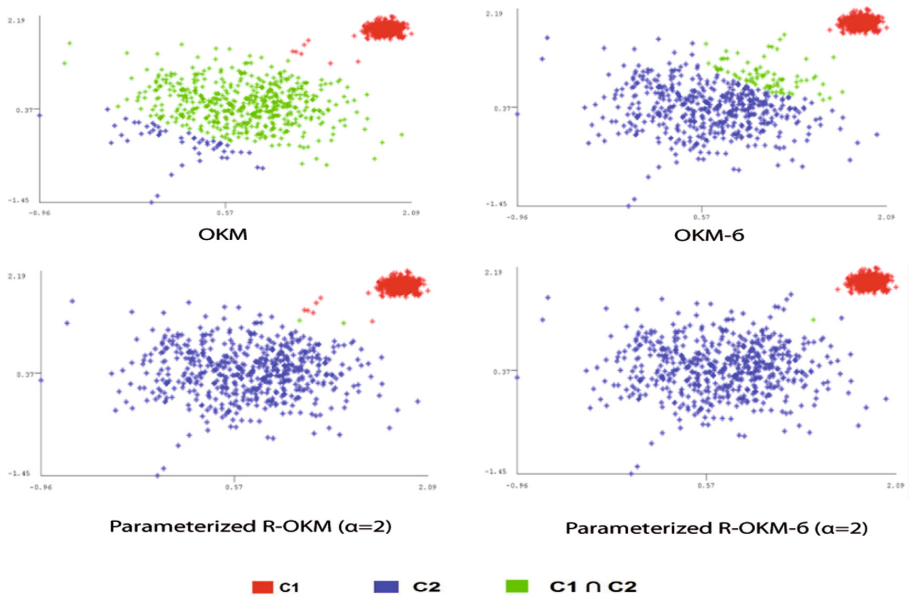


Fig. 4. Comparison of clusters obtained with OKM-σ and Parameterized R-OKM-σ versus clusters obtained with OKM and Parameterized R-OKM in "Artificial data set 2"

5.2 External Assessment

Known that most of the validity measures traditionally used for external clustering assessment become obsolete for overlapping clustering, we used external validation measures (precision, recall and F-measure) calculated using the FBcubed technique as described by Amigo *et al* [13]. These measures try to estimate whether the produced clusters were correct with respect to the underlying true categories in the data. These measures were designed to take into account special structures of overlapping clustering.

We conducted experiments on different domains that motivate the overlapping clustering researches: video classification based on the users ratings (Eachmovie[1] data set), detection of emotion in music songs (Music emotion[2] data set) and clustering natural scene image (Scene[3] data set).

Table 1 reports the average values, on ten runs, of Precision, Recall and F-measure obtained with OKM-σ and Parameterized R-OKM-σ versus OKM, and Parameterized R-OKM on artificial and real overlapping benchmarks. These results show that proposed methods outperform the original ones comparing to the overall F-measure. For example, using OKM-σ in "Artificial data set 2", the obtained F-measure increases from 0.648 to 0.899 compared to OKM. In fact,

[1] `cf.http://www.grouplens.org/node/76.`
[2] `cf.http://mlkd.csd.auth.gr/multilabel.html`
[3] `cf.http://mlkd.csd.auth.gr/multilabel.html`

Table 1. Values of Precision, recall and F-measure obtained using OKM-σ and Parameterized R-OKM-σ versus values obtained using the original methods on artificial and real overlapping benchmarks

		Precision	Recall	F-measure
Artificial data set 1	OKM	0.731	0.999	0.844
	OKM-σ	0.809	0.999	0.894
	Parameterized R-OKM	0.848	0.996	0.916
	Parameterized R-OKM-σ	0.895	0.999	**0.944**
Artificial data set 2	OKM	0.527	0.996	0.648
	OKM-σ	0.831	0.999	0.906
	Parameterized R-OKM	0.824	0.989	0.899
	Parameterized R-OKM-σ	0.898	0.999	**0.945**
Emotion data set	OKM	0.390	0.512	**0.443**
	OKM-σ	0.381	0.506	**0.435**
	Parameterized R-OKM	0.447	0.387	0.415
	Parameterized R-OKM-σ	0.395	0.448	0.419
Eachmovie data set	OKM	0.271	0.902	0.416
	OKM-σ	0.346	0.757	0.475
	Parameterized R-OKM	0.482	0.746	0.582
	Parameterized R-OKM-σ	0.500	0.751	**0.601**
Scene data set	OKM	0.145	0.938	0.251
	OKM-σ	0.193	0.996	0.324
	Parameterized R-OKM	0.425	0.391	0.407
	Parameterized R-OKM-σ	0.430	0.390	**0.412**

the improvement of F-measure in all data sets is induced by the improvement of Precision. This result is explained by the fact that proposed methods build more reduced overlaps as the difference of densities between the clusters becomes more important.

6 Conclusion

We proposed in this paper two new methods, referred to as OKM-σ and Parameterized R-OKM-σ, able to produce relevant non disjoint clusters when data contain groups with different densities. These new methods introduced a new regularization factor in the optimized criterion which regularizes the variation in densities between the obtained clusters. Experiments performed on artificial and real world overlapping benchmarks show the effectiveness of proposed methods compared to the existing ones to build more relevant clusters. This proposed methods can be applied for many other application domains where a data point needs to be assigned to more than one cluster. We plan to conduct experiments in others real and artificial overlapping data sets.

References

1. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, MULTIMEDIA 2006, pp. 421–430. ACM, New York (2006)
2. Wieczorkowska, A., Synak, P., Ras, Z.: Multi-label classification of emotions in music. In: Intelligent Information Processing and Web Mining. Advances in Soft Computing, vol. 35, pp. 307–315 (2006)
3. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
4. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8), 651–666 (2010)
5. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers Amp; Geosciences 10(23), 191–203 (1984)
6. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1, 98–110 (1993)
7. Masson, M.H., Denux, T.: Ecm: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41(4), 1384–1397 (2008)
8. Qinand, A.K., Suganthan, P.N.: Kernel neural gas algorithms with application to cluster analysis. In: Int. Conf. on Pattern Recognition, vol. 4, pp. 617–620 (2004)
9. Depril, D., Van Mechelen, I., Mirkin, B.: Algorithms for additive clustering of rectangular data tables. Computational Statistics and Data Analysis 52(11), 4923–4938 (2008)
10. Wilderjans, T.F., Depril, D., Mechelen, I.V.: Additive biclustering: A comparison of one new and two existing als algorithms. J. of Classification 30(1), 56–74 (2012)
11. Ben N'Cir, C.E., Cleuziou, G., Essoussi, N.: Identification of non-disjoint clusters with small and parameterizable overlaps. In: 2013 Int. Conf. on Computer Applications Technology (ICCAT), pp. 1–6 (2013)
12. Tsai, D.M., Lin, C.C.: Fuzzy c-means based clustering for linearly and nonlinearly separable data. Pattern Recogn. 44(8), 1750–1760 (2011)
13. Amigo, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12(4), 461–486 (2009)

Segmentation of Crop Nutrient Deficiency Using Intuitionistic Fuzzy C-Means Color Clustering Algorithm

P. Balasubramaniam* and V.P. Ananthi

Department of Mathematics,
Gandhigram Rural Institute - Deemed University,
Gandhigram - 624 302, Tamilnadu, India
balugru@gmail.com

Abstract. Nowadays, crop nutrient deficiency is common in most of the agricultural fields in India due to inadequate and imbalanced fertilization. The main aim of this work is to segment and calculate the percentage of nutrient deficiency which helps to predict the rate of fertilization needed for that crop. In this paper, a new intuitionistic fuzzy c-means color clustering algorithm (IFCM) is introduced using intuitionistic fuzzy sets (IFSs) with its distance function defined from similarity measure. Initially, all the experimental images are preprocessed. Then the preprocessed images are segmented by using the proposed clustering algorithm. The experimental results obtained by IFCM algorithm are compared with fuzzy c-means algorithm (FCM) to show the effectiveness of the proposed algorithm. Comparison results reveal that the proposed segmentation method is capable of segmenting uncertain crop images with nutrient deficiency.

1 Introduction

Evolution of deficiency and diseases in plants are horrible to today's agricultural world. Plants become victim of undesirable climate, environment, soil conditions, nutritional imbalance and so on. Being India an agricultural country, it has to increase the yield of productivity for its exponentially increasing population. There are several constraints affecting the productivity of crops in India, namely, soil type, monsoon condition, nutrient deficiency, disease, pest attack and economical base. This paper concerns with a constraint nutrient deficiency especially in crops. Nutrient deficiency in crops can be viewed by its physiological responses called symptoms and these occurs mainly in leaf and stem portion of the crop. Nutrient deficiency in crops induces a significant reduction of agricultural products. So, it is essential to monitor crops. Monitoring crops by human resource are more expansive and exhaustive. Hence, in order to improve cultivation and yield, machine supervision is needed to find diseases in crops [1]. Mao et al. [2] introduced fuzzy c-means clustering (FCM) to segment

* Corresponding author.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 112–119, 2013.
© Springer International Publishing Switzerland 2013

the disease region in crop images and the same algorithm is implemented in [3] to segment nutrient deficiency in crop images. Image analysis directly relay on segmentation. Segmentation is a procedure for classifying targets from an image. Imaging crops from their fields would be uncertain and unclear due to various environmental conditions and imaging equipments. FCM clustering is capable of removing uncertainty but not in full-fledged form. In images uncertainties occurs in terms of vagueness in imprecise gray levels, boundaries and so on. Due to the presence of uncertainties in the gray levels, hesitation arises while defining membership function of these imprecise gray levels which is also an uncertainty. Hence, fuzzy sets cannot have the ability to remove these type of uncertainty. For these reasons, Atanassov [4] introduced a higher version of fuzzy set called intuitionistic fuzzy set (IFS) in 1986 with three parameters. Chaira [5] proposed a new c-means algorithm based on Atanassov's IFSs.

This paper investigates the segmentation of nutrient deficient crop images. Initially, images in RGB color space are transformed to HSV color space to improve the quality and then the color transformed image is enhanced using median filter by removing noise. Then a new IFCM clustering algorithm whose distance function derived from the similarity measure is used to segment nutrient deficiency lesions in crop images. Experimental results are drawn to show the segmentation performance and is compared with formal FCM algorithm. Performance is evaluated quantitatively by calculating the number of iterations and time required for the segmentation process. Then percentage of deficiency is calculated from the segmented image, which helps to find the quantity of fertilizer need for that particular crop.

Frame of this paper is traced as follows. Section 2 describes the basic ideas on IFSs. Section 3 constructs IFSs using Sugeno type generator. Section 4 addresses the segmentation process using the proposed IFCM algorithm. Section 5 deals with the experimental results and discussion and finally conclusions are drawn in section 6.

2 Basic Concepts of Intuitionistic Fuzzy Sets (IFSs)

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set. A fuzzy set F of X can be mathematically expressed as $F = \{(x, \mu_F(x))|x \in X\}$, where the function $\mu_F(x) : X \to [0, 1]$ pertains the degree of membership of an element x in X.

Similarly, an IFS F in X can be written as

$$F = \{(x, \mu_F(x), \nu_F(x))|x \in X\},$$

where the functions $\mu_F(x), \nu_F(x) : X \to [0, 1]$ represent the degree of membership and non-membership of an element x in X, respectively, with the necessary restraint $0 \leq \mu_F(x) + \nu_F(x) \leq 1$. Since, this paper deals with segmentation of uncertain image so, hesitation arises in determining the function of membership owing to lack of knowledge. Atanassov introduced a third parameter $(\pi_F(x))$ called hesitation degree in IFS to eliminate these uncertainties and this degree is formulated as $\pi_F(x) = 1-\mu_F(x)-\nu_F(x)$ satisfying the inequality $0 \leq \pi_F(x) \leq 1$.

3 Generating Intuitionistic Fuzzy Sets (IFSs)

Generally, intuitionistic fuzzy generators are used to construct Atanassov's IFSs. A function $\xi(x) : [0,1] \to [0,1]$ is a continuous, increasing and decreasing generator of an IFSs if $\xi(x) \leq (1-x)$ for every $x \in [0,1]$ with $\xi(0) \leq 1$ and $\xi(1) \leq 0$, for further information see [6].

In this paper, Yager generating function is utilized to generate intuitionistic fuzzy complement (IFC) as in [7]. The function $Y(x) = x^{\gamma}$ is used to generate Yager class and this in turn used in the fuzzy complement function, defined as

$$N(\mu_F(x)) = Y^{-1}(Y(1) - Y(\mu_F(x)),$$

where $Y(\cdot)$ is an increasing function $[0,1]$ to $[0,1]$ and the membership function $\mu_F(x) = \frac{x-min(x)}{max(x)-min(x)}$, for generating Yager's IFC. Therefore, Yager's IFC is calculated as

$$N(x) = (1 - x^{\gamma})^{1/\gamma}, \quad \gamma > 0,$$

where $N(1) = 0$ and $N(0) = 1$, which is used to generate the values of non-membership degree. Thus, an IFS is constructed as

$$F_{IFS} = \{(x, \mu_F(x), (1 - \mu_F(x)^{\gamma})^{1/\gamma}) | x \in X\}$$

with the hesitation degree

$$\pi_F(x) = 1 - \mu_F(x) - (1 - \mu_F(x)^{\gamma})^{1/\gamma}. \tag{1}$$

4 Stages of Segmentation Process

Every image needs to be preprocessed before commencing the process of segmentation. Image pre-processing includes image acquisition, color transition, image enhancement and so on. In this paper, it is assumed that the images are already registered and filed in image database. After segmentation, the segmented image is to be post processed from which features are extracted to determine which nutrient is deficient. But, post processing is beyond this paper.

4.1 Color Space Conversion for Plant Images

Images in RGB space is transformed to another color space to emphasize the deficiency region. Each such transition renders a peculiar information of the image. Hence, a relevant color space is to be considered before segmentation. All the experimental images in RGB space is transformed to HSV space, since HSV model has the ability to deal with intensity variation by separating intensity values. In HSV color model, H is the parameter representing the value of hue in the range $[0°, 360°]$, S is the parameter representing saturation between the purity of the chosen color and the highest purity of the same color ranging from 0 to 1 and the third parameter V describes the brightness and varies from 0 to 1. An image I in

RGB space is converted into HSV space by using the following transformation as in [3]:

$$H = \begin{cases} \frac{60*(G-B)}{mx-min}, & R > max(G,B) \\ \frac{180*(B-R)}{mx-min}, & G > max(R,B), \\ \frac{300*(R-G)}{mx-min}, & B > max(R,G) \end{cases} \quad S = \frac{mx-min}{mx}, \quad V = mx,$$

where $H,\ S,\ V$ symbolize the parameter of hue, saturation and brightness respectively and $R,\ G,\ B$ respectively represent the red, green and blue component of the image I, further $mx = max(R,G,B)$ and $min = min(R,G,B)$.

4.2 Image Enhancement

Median filter is used to enhance the image by reducing noise. Steps for enhancing the image by using median filter are as follows:

1. Consider a 3×3 (or 5×5) region centered around the $(i,j)^{th}$ pixel.
2. Sort values of the pixel points in the selected 3×3 (or 5×5) region into ascending order.
3. Estimate the middle value in the sorted pixel which is adopted as a new value of the $(i,j)^{th}$ pixel.

It is found that a 3×3 square filter minimizes noise in a great extent than that of 5×5 square filter.

4.3 Segmentation by IFCM Algorithm

This paper utilizes a new method of clustering algorithm based on IFSs. Ability of clustering technique directly relay on the distance measure employed in its algorithm. Usually, FCM and IFCM algorithm utilize Euclidean distance as distance function. But these functions only depend on the characteristic vectors and not on the quality. In order to overcome these drawbacks, Pelekis et. al [8] introduced a new similarity measure between two IFSs F_1 and F_2, which is mathematically defined as

$$S(F_1,F_2) = \frac{s(\mu_{F_1},\mu_{F_2}) + s(\nu_{F_1},\nu_{F_2})}{2} \tag{2}$$

where

$$s(F_1^*,F_2^*) = \begin{cases} \frac{\sum_{i=1}^{n} min(F_1^*(x_i),F_2^*(x_i))}{\sum_{i=1}^{n} max(F_1^*(x_i),F_2^*(x_i))}, & F_1^* \cup F_2^* \neq \phi, \\ 1, & F_1^* \cup F_2^* = \phi, \end{cases}$$

with $F_1^*,\ F_2^* \in F_{IFS}$.

Distance function determined using Eqn. (2) is implemented in IFCM clustering algorithm to segment nutrient deficiency lesions in the crop images. Steps of the segmentation process are demonstrated below:

Step–1:

Let I be a crop image obtained after preprocessing the original source image of size $n\ (= P \times Q)$. Set the cluster class $c\ (1 < c < n)$, fuzzy index β, termination limit $\varepsilon > 0$ and iteration counter t.

Step–2:

Set the initial iteration counter $V^{(t)}$ at $t = 0$.

Step–3:

Calculate the membership matrix U for t^{th} iteration using the following equation. For all $1 \leq i \leq c,\ 1 \leq k \leq n$,

$$u_{ik}^{(t)} = \begin{cases} \left[\frac{|x_k - v_i|_{IFS}^{(t)}}{\sum_{i=1}^{c} |x_k - v_i|_{IFS}^{(t)}}\right]^{\frac{1}{1-\beta}}, & if\ S(x_k, v_i)^{(t)} \neq 1, \\ 1, & otherwise, \end{cases}$$

where $\mid x_k - v_i \mid_{IFS} = S(x_k, v_i)$.

Step–4:

Calculate a new membership matrix U^* for t^{th} iteration by using the equation $u_{ik}^{*(t)} = u_{ik}^{(t)} + \pi_{ik}^{(t)}$, where $\pi_{ik}^{(t)}$ is calculated by applying the membership values obtained in step 3 into the Eqn. (1).

Step–5:

Update cluster center $V^{(t+1)}$ by using the following equation.

$$v_i^{(t+1)} = \frac{\sum_{k=1}^{n} (u_{ik}^{*(t+1)})^{\beta} x_k}{\sum_{ik} (u_{ik}^{*(t+1)})^{\beta}}, \quad i = 1, 2, \cdots, c.$$

Step–6:

Check the distance between the membership matrix obtained from $(t+1)^{th}$ and t^{th} iterations, that is $\parallel U^{*(t+1)} - U^{*(t)} \parallel = \varepsilon$, ε is an user defined value.

It is to be noted that the new membership matrices obtained are selected in such a way to minimize the objective function J_β, defined as

$$J_\beta = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik}^*)^{\beta} \sum_{i=1}^{c} \mid x_k - v_i \mid_{IFS} + \sum_{i=1}^{c} \pi_i^* e^{1-\pi_i^*},$$

where $\pi_i^* = \frac{1}{n} \sum_{k=1}^{n} \pi_{ik},\ i = 1, 2, \cdots, c$, denote the intuitionism in fuzzy set.

5 Experimental Results and Discussion

Experimental tests are executed on 30 images and six of them are shown in figures 1(a)-1(f), where figure 1(a) shows a banana leaf with Zn deficiency, figure 1(b) paints a groundnut leaf with Mn deficiency, figures 1(c) and 1(d) portray S and Zn deficiency symptoms in sorghum crop, Mo and N deficiencies in sugarcane leaf are pictured in figure 1(e) and figure 1(f) respectively. These original

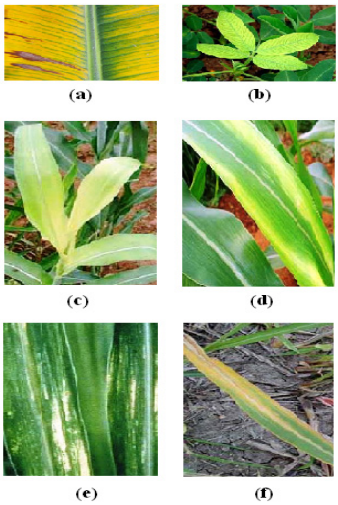


Fig. 1. Original source crop images with nutrient deficiency. (a) Banana leaf with Zn deficiency, (b) Groundnut crop with Mn deficiency, (c) Sorghum crop with S deficiency, (d) Sorghum crop with Zn deficiency, (e) Sugarcane leaves with Mo deficiency, (f) Sugarcane leaf with N deficiency.

source images are downloaded form the web portal of TamilNadu Agricultural University (TNAU).

Results of FCM clustering algorithm and the proposed method are shown in figures 2(a)-2(l). The first and third column of figure 2 show the segmented lesions obtained using FCM algorithm. Second and fourth column of figure 2 show the segmented output obtained by the proposed method. In figure 2, deficiency regions are painted in yellow color and the non-deficient regions are represented in magenta color. Figures 2(a) and 2(b) show the segmented image of figure 1(a) by FCM and the proposed IFCM respectively. Deficiency lesions of groundnut leaf segmented by FCM and the proposed method are pictured in figures 2(c) and 2(d) respectively. Similarly, segmented results of figure 1(c) obtained using FCM and the proposed algorithm are respectively shown in figures 2(e) and 2(f). Figures 2(g) and 2(h) respectively, render the lesions of the sorghum crop (shown in figure 1(d)) clustered by FCM and the proposed method. Sugarcane leaves in figures 1(e) and 1(f) are clustered by FCM and the proposed IFCM technique and their resulting images are shown in figures 2(i)-2(j) and figures 2(k)-2(l) respectively. It is clearly seen that the lesions obtained by the proposed method are clear and well segmented because it clusters relevant pixels by eliminating uncertainties from their gray levels.

Quantitative comparisons are established by finding the maximum number of iterations and time required for clustering the crop images using FCM and the proposed method. These values are given in table 1 and it shows that the proposed method needs very less time and opts less number of iteration for clustering than that of FCM technique. The edges of the resulting segmented images are super

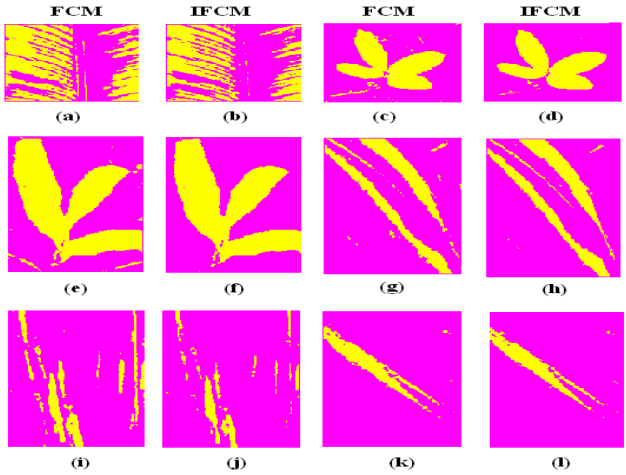


Fig. 2. (a) and (b) show the segmented lesions of figure 1(a) by using FCM and IFCM respectively, (c) and (d) show the segmented lesions of figure 1(b) by using FCM and IFCM respectively, (e) and (f) show the segmented lesions of figure 1(c) by using FCM and IFCM respectively, (g) and (h) show the segmented lesions of figure 1(d) by using FCM and IFCM respectively, (i) and (j) show the segmented lesions of figure 1(e) by using FCM and IFCM respectively, (k) and (l) show the segmented lesions of figure 1(f) by using FCM and IFCM respectively.

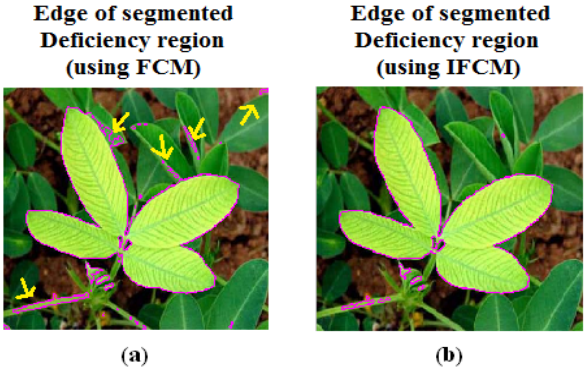


Fig. 3. Accuracy of IFCM segmentation

imposed on their original image to show the exactness of lesion segmented by IFCM algorithm. The edges of segmented images (figures 2(c) and 2(d)) superimposed on its original image (figure 1(b)) are shown in figure 3. Figure 3(a) clarifies that clustering by FCM takes pixels other than deficiency region or by leaving some portion of deficiency and figure 3(b) clustering by IFCM perfectly segments the deficiency region. The yellow arrows in figure 3, denote the irrelevant pixel region clustered excessively by FCM. Percentage of deficiency in crop images is calculated as a fraction of total deficiency region over the total area of the crop

image. Proportion of deficiency retrieved by FCM and the proposed method are given in table 1. From the table it is obviously seen that the deficiency percentage obtained by FCM is redundant than the proposed method because it clustered some of the irrelevant pixels than deficiency. Hence, from the results it is clear that the proposed method segments well by capturing uncertainties.

Table 1. Performance comparison

Nutrient deficient	*Time taken (sec)*		*No. of iteration*		*Deficiency(%)*	
crop images	*FCM*	*IFCM*	*FCM*	*IFCM*	*FCM*	*IFCM*
Banana (Zn def)	2.7962	1.5422	29	9	50.10	42.87
Groundnut (Mn def)	2.7992	1.8944	14	6	10.13	9.84
Sorghum (S def)	2.8310	1.4434	21	6	57.05	55.17
Sorghum (Zn def)	12.6569	3.4834	56	11	20.49	20.27
Sugarcane (Mo def)	4.7565	2.8723	24	11	11.37	7.72
Sugarcane (N def)	12.7261	4.0553	95	23	16.29	15.10

6 Conclusion

This paper renders a new approach for nutrient deficiency segmentation of crop images based on IFSs. Most of the segmentation algorithm depends on prior knowledge. IFCM eliminates this complexity by using hesitation degree. The proposed technique is tested on several crop images and the results reveals that the segmented images obtained are clear and properly clustered than FCM method. Quantitatively, the proposed way of clustering is far better than FCM algorithm. Membership function defined is not always precise due to hesitation and IFSs render better results by eliminating such vagueness and rate of fertilization can be assessed using the percentage of deficiency computed from the crops.

References

1. Camargo, A., Smith, J.S.: An image processing based algorithm to automatically identify plant disease visual symptoms. Biosyst. Eng. 102, 9–21 (2009)
2. Mao, H.P., Zhang, Y.C., Hu, B.: Segmentation of crop disease leaf images using fuzzy c-means clustering algorithm. Trans. Chin. Soc. Agric. Eng. 24, 136–140 (2008)
3. Hu, J., Li, D., Chen, G., Duan, Q., Han, Y.: Image segmentation method for crop nutrient deficiency based on fuzzy c-means algorithm. Intell. Autom. Soft Comput. 18, 1145–1155 (2012)
4. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets and Syst. 20, 87–96 (1986)
5. Chaira, T.: A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images. Appl. Soft Comput. 11, 1711–1717 (2011)
6. Bustince, H., Kacpryzk, J., Mohedano, Z.: Intuitionistic fuzzy generators application to intuitionistic fuzzy complementation. Fuzzy Sets and Syst. 114, 485–504 (2000)
7. Burillo, P., Bustince, H.: Entropy on intuitionistic fuzzy sets and on interval- valued fuzzy set. Fuzzy Sets ans Syst. 78, 305–316 (1996)
8. Pelekis, N., Iakovidis, D.K., Kotsifakos, E.E., Kopanakis, I.: Fuzzy clustering of intuitionistic fuzzy data. Int. J. Bus. Intell. Data Min. 3, 45–65 (2007)

An Efficient Artificial Bee Colony and Fuzzy C Means Based Co-regulated Biclustering from Gene Expression Data

K. Sathishkumar[1], E. Balamurugan[2], and P. Narendran[1]

[1] Gobi Arts & Science College, Gobichettipalayam
[2] Bannari Amman institute of Technology, Sathyamangalam
{Sathishmsc.vlp,rethinbs,narendranp}@gmail.com

Abstract. The gene microarray data are arranged based on the pattern of gene expression using various clustering algorithms and the dynamic natures of biological processes are generally unnoticed by the traditional clustering algorithms. To overcome the problems in gene expression analysis, novel algorithms for finding the coregulated clusters, dimensionality reduction and clustering have been proposed. The coregulated clusters are determined using biclustering algorithm, so it is called as coregulated biclusters. The coregulated biclusters are two or more genes which contain similarity features. The dimensionality reduction of microarray gene expression data is carried out using Locality Sensitive Discriminant Analysis (LSDA). To maintain bond between the neighborhoods in locality, LSDA is used and an efficient meta heuristic optimization algorithm called Artificial Bee Colony (ABC) using Fuzzy C Means clustering is used for clustering the gene expression based on the pattern. The experimental results shows that proposed algorithm achieve a higher clustering accuracy and takes lesser less clustering time when compared with existing algorithms.

Keywords: Gene expression data, Bimax Algorithm, Co-regulated Biclusters, Locality Sensitive Discriminant Analysis, Artificial Bee Colony, Fuzzy C Means.

1 Introduction

The purpose of clustering gene expression data is to reveal the natural structure inherent in the data. A good clustering algorithm should depend as little as possible on prior knowledge, for example requiring the predetermined number of clusters as an input parameter. Clustering algorithms for gene expression data should be capable of extracting useful information from noisy data. Gene expression data are often highly connected and may have intersecting and embedded patterns [1,2]. Therefore, algorithms for gene-based clustering should be able to handle this situation effectively. Finally, biologists are not only interested in the clusters of genes, but also in the relationships (i.e., closeness) among the clusters and their sub-clusters, and the relationship among the genes within a cluster (e.g., which gene can be considered as

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 120–129, 2013.
© Springer International Publishing Switzerland 2013

the representative of the cluster and which genes are at the boundary area of the cluster) [3, 4].

2 Related Works

K-means is a typical partition-based clustering algorithm used for clustering gene expression data. It divides the data into pre-defined number of clusters in order to optimize a predefined criterion. The major advantages of it are its simplicity and speed, which allows it to run on large datasets [5]. However, it may not yield the same result with each run of the algorithm. Often, it can be found incapable of handling outliers and is not suitable to detect clusters of arbitrary shapes. Self Organizing Map (SOM) is more robust than K-means for clustering noisy data. It requires the number of clusters and the grid layout of the neuron map as user input. Specifying the number of clusters in advance is difficult in case of gene expression data [6].

K-nearest neighbor based density estimation technique is proposed [7]. Another density based algorithm proposed by Chung et al. works in three phases: density estimation for each gene, rough clustering using core genes and cluster refinement using border genes. A density and shared nearest neighbor based clustering method is presented [8]. The similarity measure used is that of Pearson's correlation and the density of a gene is given by the sum of its similarities with its neighbors. The use of shared nearest neighbor measure is justified by the fact that the presence of shared neighbors between two dense genes means that the density around the dense genes is similar and hence should be included in the same cluster along with their neighbors.

Fuzzy C-means (FCM) is an extension of K-means clustering and bases the fuzzy assignment of an object to a cluster on the relative distance between the object and all cluster centroids. Many variants of FCM have been proposed in the past years, including a fuzzy clustering approach, FLAME [9], which detects dataset-specific structures by defining neighborhood relations and then neighborhood approximation of fuzzy memberships are used so that non-globular and nonlinear clusters are also captured.

3 Methodology

The proposed approach consists of three stages namely finding of coregulated biclusters using Bimax algorithm, dimensionality reduction using Locality Sensitive Discriminant Analysis (LSDA) and clustering using MoABC.

3.1 Finding of Coregulated Clusters Using Bimax Algorithm

The diagram for finding the coregulated clusters using algorithm is shown in Fig.1. Enhanced Bimax algorithm is used to display a maximal biclusters value and displays

a coregulated biclusters. The Enhanced Bimax algorithm is used to measure a particular gene is present or not. It also finds the transcription sites of the coregulated biclusters. Normalization technique used to specify genes are presented in the particular group or not. The output is display the transcription factors.

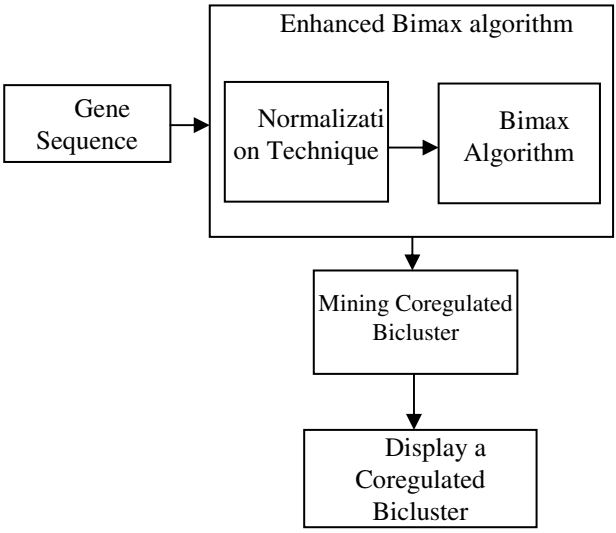


Fig. 1. Block diagram for mining coregulated bicluster

3.2 Bimax Algorithm

The Bimax algorithm needs to guarantee that only optimal, inclusion-maximal biclusters are generated. It is used to specify the genes and conditions. It is used to specify that analysis of DNA chips and gene networks. The algorithm realizes the divide-and-conquer strategy. Fig. 1 describes an original Bimax algorithm. It consists of three procedures. They are Enhanced Bimax. Conquer and Divide. Conquer function is call and check the condition is if the genes and conditions are equal then the partitioning is begin, otherwise it stop the process. Second step is split the data and normalization technique is used to group the splited data. It is used to find all add the maximum groups in general gene expression data. Each coregulated genes are grouping together the particular expression value and the particular situation [13].

3.3 Proposed Enhanced Bimax Algorithm

Enhanced Bimax algorithm can contain two procedures. Fig. 2 describes a flowchart for proposed Enhanced Bimax algorithm.

Binary Space Partitioning (BSP) is a method for recursively subdividing a space into convex sets by hyper planes. This subdivision gives rise to a representation of the scene by means of a tree data structure known as a BSP tree. Normalization is the process of isolating statistical error in repeated measured data. Quintile normalization for instance, is normalization based on the magnitude of the measures. The goals in doing eliminate all the redundant data and ensure data dependencies. The numbers of genes that reproducibly showed and the unnormalized data and normalized data are displayed on the coregulated biclusters. Enhanced Bimax algorithm is applied data mining technique on clustering. In the clustering similar samples and similar gene probes are organized in a fashion so that they would lie close together. It consists of three procedures. They are Enhanced Bimax, Breadth-First Search (BFS) and BSP [13].

They are BFS and BPS combination of sequences searches the entire graph c nodes of a graph or rods, it exhaustively y Space Partitioning. First step is normalization technique used to remove the redundant data and then grouping genes in the specific conditions. Binary Space Partitioning function is call and check the condition is if the genes and conditions are equal then the partitioning is begin. Otherwise it stop the process. It specifies that a particular gene is present in the given group then it is represents a one. With these maximum groups in general gene expression data can be found. Otherwise the gene is not present in the given group then it is representing as zero. Fig. 2 describes a proposed Enhanced Bimax algorithm.

3.4 Locality Sensitive Discriminant Objective Function for Dimensionality Reduction.

It is observed that naturally occurring data may be generated by structured systems with possibly much fewer degrees of freedom than the ambient dimension would suggest, a number of research works have been developed with the case considering when the data lives on or close to a submanifold of the ambient space [10].

3.5 Proposed Artificial Bees Colony Based Fuzzy Clustering

The modifications carried out to improve the basic ABC algorithm and its application used to achieve fuzzy clustering is been given in this section.

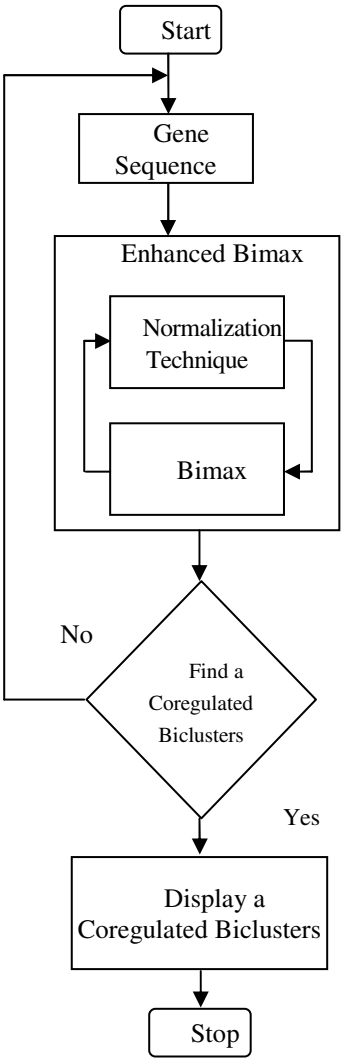


Fig. 2. Flow Chart for Proposed Enhanced Bimax Algorithm

3.6 Locality Sensitive Discriminant Analysis

A novel linear dimensionality reduction algorithm called Locality Sensitive Discriminant Analysis (LSDA). For the class of spectrally based dimensionality reduction techniques, it optimizes a fundamentally different criterion compared to classical dimensionality reduction approaches based on Fisher's criterion (LDA) or Principal Component Analysis.

3.7 Fuzzy C-Means Clustering (FCM)

FCM is a clustering algorithm which allows one data may belong to two or more clusters. It is normally used in pattern recognition [11]. It is based on minimization of the following objective function (1):

$$j_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, 1 \le m < \infty \quad (1)$$

Where,

$m =$ is any real number greater than 1

$u_{ij} =$ is the degree of membership of xi in the cluster j

x_i is the ith of d-dimensional data

c_j is the cluster centre of d-dimension data

$\left\|x_i - c_j\right\|^2$ is the distance measured of similarity between the measured data and the cluster data

3.8 Artificial Bee Colony Algorithm

It is a swarm intelligent method which inspired from the intelligent foraging behavior of honey bee swarms. It strength is its robustness and its simplicity. It is developed by surveying the behavior of the bees is finding the food source which is called nectar and sharing the information of food source the bee which is present in the nest. In the ABC the artificial agents are classify into three types; such as employed bee, the onlooker bee and the scout each of the bee plays different role in the process. The employed bee stays on a food source and in its memory provides the neighborhood of the food source. Each employed bee carries with her information about the food source and shares the information to onlooker bee. The onlooker bees wait in the hive on the dance area, after getting the information from employed bees about the possible food source then make decision to choose a food source in order to use it. The onlooker bees select the food source according to the probability of that food source. The food source with lower quantity of nectar that attracts less onlooker bees compared to ones with a higher quantity of nectar. Scout bees are searching randomly for a new solution. The employed bee whose food source has been abandoned it becomes a scout bee. The goal of the bees in the ABC model is to find the best solution. In the ABC algorithm the number of employed bees is equal to the number of onlooker bees which is also equal to the number of solutions. The ABC algorithm consist of a Maximum Cycle Number (MCN) during each cycle, there are three main parts:

- Sending the employed bees to the food sources and calculate their nectar quantities
- Selecting the food sources by the onlooker bees
- Determining the scout bee and discover a new possible food sources

Employed Bee

In the employed bee phase, each employed bee determines a new solution from the neighborhood of the current food source.

The employed bee compared the current solution with the new solution and memorizes the best one by apply the greedy selection process. When all employed bees have finished this search process, then they share the fitness value (nectar information) and the position of the food source (solution) to the onlooker bees.

Onlooker Bee

In the onlooker bee phase, after getting the information about the nectar and position of the food source each onlooker bee selects a food source with a probability of higher nectar information.

Scout Bee

If a food source position cannot be improved through fixed cycles, it is called 'limit', it means that the solution has been sufficiently exploited, and it may be removed from the population.

MoABC Based FCM

In order to perform fuzzy clustering for image segmentation using the proposed MoABC-FCM algorithm, a population of SN $(z_1, z_2, z_3 \ldots . z_{SN})$ solutions is created, where SN is the number of employed bees or onlooker bees. Each bee represents a potential solution of the fuzzy clustering problem. Each individual bee z_i in generation G is formulated using equation (2):

$$z_i(G) = (v_{i,1}, v_{i,2}, \ldots \ldots v_{i,c})^T, \ subject\ to\ 1 \leq i \leq SN \tag{2}$$

C is the number of clusters and, $v_{i,k}$ represents the kth cluster center for the ith bee.

The goal of MoABC-FCM algorithm is to determine when algorithm gets into convergence; it is converted into the optimal fuzzy partition matrix to a crisp partition matrix. The defuzzification is carried out by assigning each pixel to the cluster with the highest membership.

4 Results and Discussion

The proposed technique for microarray gene clustering has been implemented in the working platform of MATLAB (version 7.11). For evaluating the proposed technique, the microarray gene samples of human acute leukemia and colon cancer data are

utilized. [12] The high dimensional gene expression data has been subjected to dimensionality reduction and so a dimensionality reduced gene data with dimensions has been obtained. Thus LSDA is applied to identify informative genes and reduce gene dimensionality for clustering samples to detect their phenotypes.

Table 1. Microarray gene data dimension utilized for the evaluation process

Types of Gene Data	Number of Samples	Number of Genes	Dimensionality Reduced Data with the aid of LPP
ALL	41	7139	41X41
AML	36	7128	36X40
COLON	68	3000	62X42

A sample of microarray gene dataset of three classes that has been used for testing is given in the Table 2. Clustering for microarray gene expression data whose amount is large can be fully calculated by determining the boundary of the clusters.

Table 2. A sample of the microarray gene data to test the proposed technique

Class	ALL		AML		COLON	
Sample gene	ALL 16125 TA-Norel	ALL 23668 TA-Norel	AML SH-5	AML SH-13	AFFX-MurIL2	AFFX-MurIL10
AFFX-CreX-5_at (endogenous control)	-172A	- 93A	-271A	-11A	20.6	-16
AFFX-CreX-3_at (endogenous control)	52A	10A	-12A	112A	-8.7	41.2
AFFX-BioB-5_st (endogenous control)	-134A	159A	-104A	-176A	4880	26.2

While testing, when a gene dataset is given, the proposed technique has to identify its belonging cluster. Existing clustering algorithms, such as Fuzzy C-means and Fuzzy Possibilistic C-Means Algorithm using EM Algorithm approaches and also MoABC are applied both to group genes, to partition samples in the early stage and have proven to be useful. The performance of each clustering algorithm may vary greatly with different data sets. Complete-link clustering method uses the smallest similarity within a cluster as the cluster similarity, and every data object within the cluster is related to every other with at least the similarity of the cluster. In order to test the performance of the data, N artificial m-dimensional feature vectors from a multivariate normal distribution having different parameters and densities were generated. Situations of large variability of cluster shapes, densities, and number of data points in each cluster were simulated.

Table 3. Performance comparison in percentage between the proposed MoABC clustering technique and other existing Techniques

Type of Gene Data	Accuracy				Correlation				Distance				Error Rate			
	FCM	FPCM	EMFPCM	MoABC	FCM	FPCM	EMFPCM	MoABC	FCM	FPCM	EMFPCM	MoABC	FCM	FPCM	EMFPCM	MoABC
ALL	83.1	83.9	85.69	87.25	0.345	0.368	0.412	0.4852	0.00379	0.00346	0.00263	0.00142	0.21	0.20	0.18	0.12
AML	80.06	81.02	83.84	85.12	0.024	0.029	0.0315	0.0396	0.00364	0.00331	0.00201	0.00185	0.30	0.29	0.24	0.16
COLON	79.0	79.9	81.96	83.04	0.119	0.125	0.139	0.215	0.02029	0.02011	0.0126	0.0099	0.04	0.03	0.01	0.006

From the Table 3, it can be seen that the proposed technique MoABC has provided more accuracy, correlation and less distance and error rate rather than the other gene clustering techniques like FCM, FPCM etc. More accuracy and less error rate leads to effective clustering of the given microarray gene data to the actual class of the gene.

5 Conclusion

Genes involved in multiple biological processes (simultaneously) may play a major role in one process while playing a minor role in another process. The importance of a gene in multiple processes has potential for further investigation. . In this paper, an effective microarray gene data clustering technique has been proposed with the aid of Bimax algorithm, LSDA and MoABC. Initially, the micro array data genes are given to the Bimax algorithm to find coregulated biclusters to reduce the space complexity of the genes, and then the dimensionality of the microarray data has been reduced with the help of LSDA mechanism. The technique has been tested by clustering the microarray gene expression data of human acute leukemia and colon cancer data. From the results, it can be noticed that our approach yields equally good results for the entire functional category. The comparative results have shown that the proposed technique possesses better accuracy, correlation and lesser distance, error rate than FCM, FPCM gene clustering techniques. The experimental results show that proposed algorithm achieved better improvement in the quality of the results by using MoABC. Hence, this means of gene clustering have paved the way for effective information retrieval in the microarray gene expression data.

References

1. Belcastro, V., Gregoretti, F., Siciliano, V., Santoro, M., D'Angelo, G., Oliva, G., di Bernardo, D.: Reverse Engineering and Analysis of Genome-Wide Gene Regulatory Networks from Gene Expression Profiles Using High-Performance Computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(3), 668–678 (2012)
2. Yuan, Y., Li, C.-T.: Partial Mixture Model for Tight Clustering in Exploratory Gene Expression Analysis. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007 (2007)
3. Yin, L., Huang, C.-H.: Clustering of Gene Expression Data: Performance and Similarity Analysis. In: First International Multi-Symposiums on Computer and Computational Sciences, IMSCCS 2006 (2006)
4. Jiang, D., Pei, J., Zhang, A.: "DHC: a density-based hierarchical clustering method for time series gene expression data'. In: Proceedings of the 3rd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, USA, p. 393 (2003)
5. Dhiraj, K., Rath, S.K., Pandey, A.: Gene Expression Analysis Using Clustering. In: 3rd International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2009 (2009)
6. Yano, N., Kotani, M.: Clustering gene expression data using self-organizing maps and k-means clustering. In: SICE 2003 Annual Conference, vol. 3, pp. 3211–3215 (2003)
7. Chung, S., Jun, J., McLeod, D.: Mining geneexpression datasets using density based clustering. Technical Report, USC/IMSC, University of Southern California, No. IMSC-04-002 (2004)
8. Syamala, R., Abidin, T., Perrizo, W.: Clustering Microarray Data based on Density and Shared Nearest Neighbor Measure. In: Proceedings of the 21st ISCA International Conference on Computers and Their Applications (CATA 2006), pp. 23–25 (2006)
9. Fu, L., Medico, E.: FLAME: A novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 8(3) (2007)
10. Cai, D., He, X., Zhou, K., Han, J., Bao, H.: Locality Sensitive Discriminant Analysis (2007)
11. Geng, X., Tao, F.: GNRFCM: A new fuzzy clustering algorithm and its application. In: International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII (2012)
12. Wen, J.: Ontology Based Clustering for Improving Genomic IR. Twentieth IEEE International Symposium International Journal of Data Mining and Bioinformatics 3(3), 229–259 (2009)
13. Chandran, C.P., IswaryaLakshmi, K.: Biclustering analysis of coregulatedbiclusters from gene expression data. International Journal of Computational Intelligence and Informatics 2(1) (2012)

Bisecting K-Means Initialization Technique for Protein Sequence Motif Identification

M. Chitralegha and K. Thangavel

Department of Computer Science, Periyar University, Salem, India

Abstract. Bioinformatics deals with the information technology as applied to management and analysis of biological data. In the field of bioinformatics, data mining helps researchers to mine large amount of biomolecular data. Major research efforts done in the area of bioinformatics involves sequence analysis, protein structure prediction and gene finding. Proteins are said to be an important molecule in all living organisms. They involve virtually in all cell functions. Protein sequence motifs are short fragments of conserved amino acids that transcend in protein sequences. Identifying such motifs is one of the challenging tasks in the area of bioinformatics. Data mining is one such technique to explore sequence motifs. These protein motifs are identified from the segments of protein sequences. All generated sequence segments may not be significant to find sequence motifs. The generated sequence segments have no classes or labels. Hence, Singular Value Decomposition (SVD) entropy technique is adopted as preprocessing method to select sequence segments. The Adaptive Fuzzy C-Means clustering method is performed on the selected segments to obtain granules. Then Bisecting K-Means is applied on each granule to obtain the specified number of clusters. These cluster centroids are given as input to the K-Means algorithm to cluster each granule separately. The result obtained using new initialization technique is then compared with random initialization for K-Means clustering. The comparative results show that new seed selection technique performs better than random initialization. This proposed method identifies significant motif patterns.

Keywords: Sequence Motifs, HSSP, SVD, Bisecting K-Means, Adaptive Fuzzy C-Means.

1 Introduction

Proteins can be considered as one of the most important elements in the process of life. They regulate variety of activities in all known organisms, from replication of the genetic code to transporting oxygen, and are generally responsible for regulating the cellular machinery and determining the phenotype of an organism. The term 'Motif' refers to a region or portion of a protein sequence that has specific structure and is functionally significant. Detection of such motifs in proteins is an important problem in today's bioinformatics research. These motif patterns may able to predict other protein's structural or functional area, such as De-oxyribo Nucleic Acid (DNA) or Ribo Nucleic Acid (RNA) binding sites, conserved domains etc.

There are several popular motif databases. PROSITE [6], PRINTS [1] and BLOCKS [5] are the three most popular motif databases. The most important motif finding tools are MITRA, Profile Branching, EMOTIF, CoSMos and Motif Scan [3].

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 130–138, 2013.
© Springer International Publishing Switzerland 2013

But, these methods will generate motif patterns only for a single protein sequence. The patterns obtained by using above methods, may carry only a little information about conserved sequence regions which transcend protein families. Instead, in this paper, a huge number of segments are generated from HSSP file [8] for all protein sequences. In this paper, variant of K-Means called Bisecting K-Means clustering is used to generate seeds for K-Means clustering. Bisecting K-Means is said to be more efficient and are less susceptible to initialization problems. Careful seed selection helps us to locate hidden sequence motifs that transcend protein sequences. These identified motifs may have biological importance in day to day life.

The rest of the paper is organized as follows. Section 2 shows related work in this area of research. Clustering algorithm is explained in section 3. In section 4 the method adopted in the proposed work for seed selection process have been explained. In section 5, experimental analysis is provided. Section 6 concludes the paper with directions for further enhancement.

2 Related Work

Han and Baker [10] have first used K-Means clustering algorithm for finding protein sequence motif. Selecting initial points randomly leads to an unsatisfactory partition because some initial points may lie close to each other. In order to overcome the above mentioned problem, Wei Zhong [10] has proposed Improved K-Means clustering to explore sequence motifs. Improved K-Means algorithm tries to obtain initial seeds by using Greedy approach. In this area of research, data set is said to be huge and selecting initial seeds using above greedy approach leads to high computational cost. Computational cost is a major problem to be faced when input data-set is very large. Hence, Bernard Chen [3] has proposed granular computing model using Fuzzy clustering technique. Fuzzy C-Means granular computing approach suffers from several constraints that affect the performance of final clustering. The drawback is that, the membership of a data point in a cluster depends directly on the membership values in other cluster centers and it sometimes produces unrealistic results. Hence in this paper Singular Value Decomposition segment selection technique is used to select significant segments [2] then Adaptive Fuzzy C-Means granular computing techniques have been adopted to obtain hidden protein sequence motifs present across different protein families.

3 Clustering Algorithm

3.1 K-Means Clustering

This section explains the original K-Means clustering algorithm. The idea is to classify a set of input samples into K number of disjoint clusters, where the value of K is fixed in advance. The algorithm consists of two separate phases: the first phase is to define K seeds, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the seeds are included in some cluster, first step is completed and initial grouping is done.

Next, we need to recalculate the new centroids by including new seeds which leads to a change in the cluster centroids. The loop continues until a situation will be

reached where centroids do not move anymore. This signifies the convergence criterion for clustering [3]. Pseudocode for the K-Means clustering algorithm is listed as Fig. 1. The K-Means algorithm is the most widely studied clustering algorithm and is generally effective in producing good results. The major drawback of this algorithm is that it produces different clusters for different set initial centroids. Therefore our proposed work tries to find potential seeds in a different manner.

Algorithm: K-Means
Input: D = {d_1, d_2,...,d_n} // set of n data points
K // Number of desired clusters
Output: Partitions of the input data into K clusters
Steps:
1: Arbitrarily choose K objects as the initial centroid
2: Repeat
3: (re)assign each remaining object to the cluster to which the object is the most Similar, based on the mean value of the object in the cluster
4: update the cluster mean
5: Until no change

Fig. 1. K-Means Clustering Algorithm

3.2 Adaptive Fuzzy C-Means

Many of the behavioural problems with the standard Fuzzy C-Means algorithm are eliminated when we relax probabilistic constraint imposed on membership function [3]. Krishnapuram and Keller [4] modified the approach for calculating membership values. In fuzzy adaptive clustering the constraint on data point memberships is imposed by Eq. 1

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \mu_j(seg_i) = n \tag{1}$$

where, $\mu_{ij}$is the membership of seg_i in j^{th} cluster, k is the specified number of clusters and n is the number of sequence segments in the database.

In adaptive fuzzy, the total membership quantifiers for all sample points are equal to n. This flexible approach leads to clustering optimization problem, provides a way to improve cluster robustness. It is in this sense the algorithm is adaptive; that is membership is based on sample size rather than fixed to upper limit such as one in Fuzzy C-Means clustering. The adaptive fuzzy clustering algorithm is efficient in handling data with outlier points. It gives very low membership values for outliers, since the sum of distances of points in all the clusters involves in membership calculation.

4 Proposed Work

In this proposed work Adaptive Fuzzy C-Means granular computing method is adopted to generate small information granules. Then on each granule, K-Means algorithm has been applied. The initial seeds for K-Means clustering are obtained using bisecting K-Means algorithm.

Algorithm: Bisecting K-Means
Input: $D = \{d_1, d_2, \ldots, d_n\}$ // set of n data points
K // Number of desired clusters
Output: Partitions of the input data into K clusters
Steps:
1: Initialize the list of clusters to contain the cluster consisting of all points.
2: Repeat
3: Remove a cluster from the list of clusters.
4: {Perform several "trial" bisections of the chosen cluster.}
5: For i = 1 to number of trials do
6: Bisect the selected cluster using basic K-Means.
7: End for
8: Select the two clusters from the bisection whose (structural similarity is >Thold1 and size of cluster < Thold2)
9: Add these two clusters to the list of clusters.
10: Until the list of clusters contain K clusters.

Fig. 2. Bi-Secting K-Means Clustering Algorithm

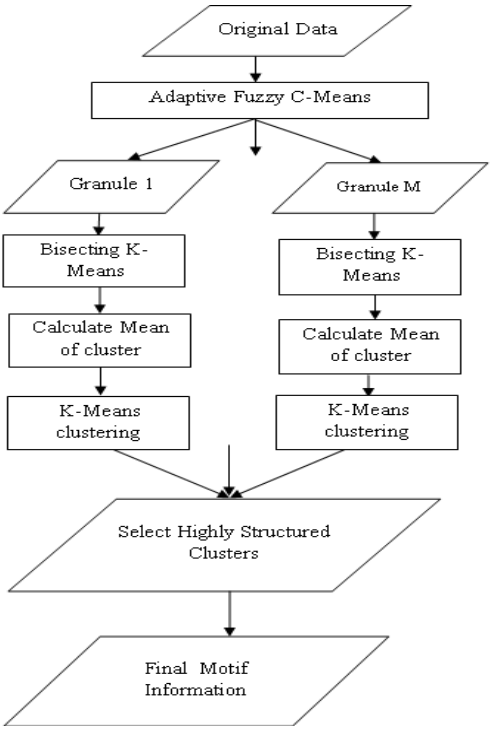


Fig. 3. Structure of Adaptive FCM Granularization with Bisecting K-Means

The basic idea behind bisecting K-Means is to split set of all points into two clusters, select one of these clusters to split based on threshold, and so on, until K clusters have been produced [9]. The mean of each cluster produced by bisecting K-Means algorithm has been used as initial seeds for K-Means. Fig. 2 shows the details of Bisecting K-Means clustering. Fig.3 shows the structure of Adaptive Fuzzy C-Means using Bi-Secting initialization technique.

5 Experimental Results

5.1 Data Set

The latest dataset obtained from Protein Culling Server (PISCES) [11] includes 4946 protein sequences. In this work, we have considered 3000 protein sequences to extract sequence motifs that transcend in protein sequences. The threshold for percentage identity cut-off is set as less than or equal to 25%, resolution cut-off is 0.0 to 2.2, R-factor cut-off is 1.0 and length of each sequence varies from 40 to 10,000.

The sliding windows with ten successive residues are generated from protein sequences. Each window represents one sequence segment of ten continuous positions. Around 6, 60, 364 sequence segments are generated by sliding window method, from 3000 protein sequences. Each sequence segment is represented by 10 X 20 matrix, where ten rows represent each position of sliding window and 20 columns represent 20 amino acids. Homology Secondary Structure Prediction (HSSP) frequency profiles are used to represent each segment [9]. Database of Secondary Structure Prediction (DSSP) assigns secondary structure to eight different classes [7]. In this paper, we convert those eight classes to three different classes based on the CASP experiment as follows [3]: H, G and I to H (Helices); B and E to E (Sheets); all others to C (Coils).

5.2 Structural Similarity

Average structural similarity of a cluster is calculated using the following formula [3]:

$$\frac{\sum_{i=1}^{ws} \text{Max}\,(P_{i,H}, P_{i,E}, P_{i,C})}{ws} \tag{2}$$

where ws is the window size and $P_{i,H}$, $P_{i,E}$ and $P_{i,C}$ shows frequency of Helices, Sheets and Coils among the segments for the cluster in position i. If the structural homology for a cluster exceeds 70% the cluster can be considered more structurally similar and if it is between 60% and 70% then the cluster is said to weakly structurally homologous.

5.3 Distance Measure

Dissimilarity between each sequence segment is calculated using city block metric. In this field of research city block metric is more suitable than Euclidean metric because it considers every position of the frequency profile equally. The following formula is used for distance calculation [3]:

$$\text{Distance} = \sum_{i=1}^{ws} \sum_{j=1}^{N} |D_s(i,j) - D_c(i,j)| \tag{3}$$

where ws is the window size and N is 20 amino acids. D_s(i, j) is the value of the matrix at row i and column j which represents sequence segment. D_c (i, j) is the value of the matrix at row i and column j which represents the centroid of a given cluster.

5.4 Davis-Bouldin Index Measure

Davis-Bouldin Index, measures how compact and well separated the clusters are. Small values of DB are indicative of the presence of compact and well separated clusters. The lower DBI value indicates the efficiency of the method used for clustering. In this paper, quality of clusters is measured using DBI index measure.

DBI index is based on similarity measure of cluster (R_{ij}) whose bases are the dispersion measure of a cluster i is (s_i) and dissimilarity measure (d_{ij}). The similarity measure of clusters (R_{ij}) be defined as [3],

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{4}$$

Subject to,

- $R_{ij} \geq 0, R_{ij} = R_{ji}$
- If $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
- If $s_i > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
- If $s_i = s_k$ then $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$

where $d_{ij} = d(z_i, z_j)$ and $s_i = \frac{1}{\|C_i\|} \sum_{x \in C_i} d(x, z_i)$

Then the Davies-Bouldin is defined as,

$$\text{DB} = \frac{1}{k} \left[\sum_{i=1}^{k} R_i\right]$$

where $R_i = \max_{j=1,2,\ldots,k, i \neq j}(R_{ij}), i = 1,2, \ldots, k$

5.5 HSSP-BLOSUM62 Measure

HSSP frequency profile and BLOSUM62 matrix has been combined to obtain significance of motif information. Hence, the measure is defined as the following [3].

If m = 0: HSSP-BLOSUM62 measure = 0
Else If m = 1:HSSP-BLOSUM62 measure = BLOSUM62_{ii}
Else: HSSP-BLOSUM62 measure =

$$\frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} HSSP_i \,.\, HSSP_j \,.\, BLOSUM62_{ij}}{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} HSSP_i \,.\, HSSP_j}$$

where m is the number of amino acids with frequency higher than certain threshold in the same position. HSSP_i indicates the percent of amino acid i to be appeared. BLOSUM62_{ij} denotes the value of BLOSUM62 on amino acid i and j. The higher HSSP-BLOSUM62 value indicates more significant motif information.

5.6 Experimental Results

It is observed from table 1 that in both cases before and after segment selection process the Adaptive FCM with K-Means using bisecting K-Means (BKM) initialization technique identifies more number of hidden motif patterns by looking towards structural similarity values. From table 1, decreased DBI value indicates the improvement of cluster quality in Adaptive FCM with K-Means using BKM initialization algorithm. The motif information obtained from proposed method is more significant than using random initialization technique by viewing the results of HSSP-BLOSUM62 measure.

Fig. 4 has been interpreted from table 1. From the above fig. 4 we state that the number of strong and weak clusters has been increased.

Table 1. Shows the comparative results obtained from different algorithms

	Before SVD Segment Selection Process		After SVD Segment Selection Process	
	AFCM using random initialization	AFCM using BKM initialization	AFCM using random initialization	AFCM using BKM initialization
No of clusters >70%St. Similarity	170	183	197	219
No of clusters > 60% < 70% St. Similarity	271	327	234	274
% of Seq. seg > 70% St. Similarity	26.59	28.98	33.15	36.13
% of Seq. Seg > 60% <70% St. Similarity	30.27	31.67	22.11	24.54
DBI Measure	3.92	3.94	3.97	3.99
Avg HSSP-BLOSUM62	0.73	0.83	0.73	0.74

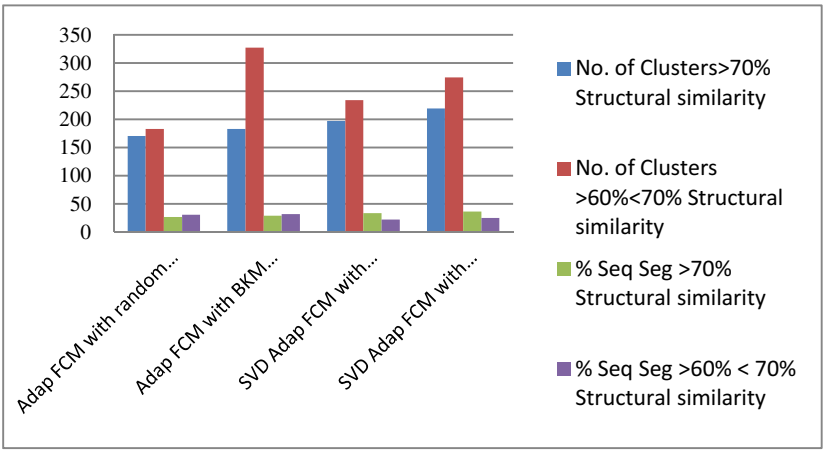


Fig. 4. Comparison of Structural Similarities

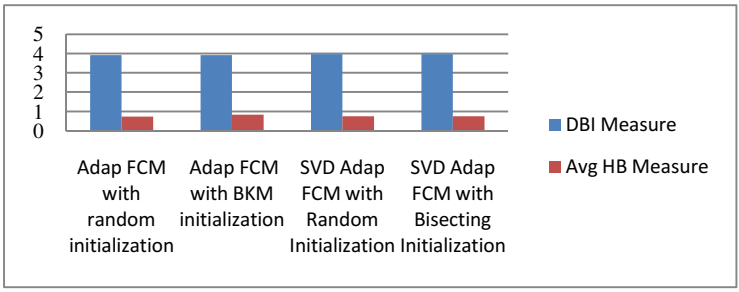
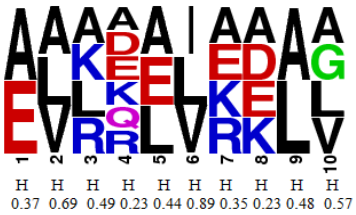
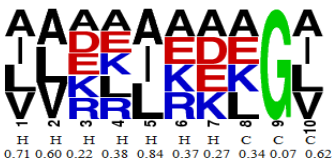


Fig. 5. Comparison of DBI and HSSP-BLOSUM62 Measure

From Fig. 4 percentage of sequence segments has also been increased in Adaptive FCM with K-Means using BKM initialization technique. Fig.5 shows comparative analysis of cluster quality and quality of motif information. Decreased DBI value and increased HSSP-BLOSUM62 values show the performance of clustering and significance of motif information obtained using proposed method.

5.7 Sequence Motifs

Two different motif patterns obtained from adaptive fuzzy C-Means granularization method are shown in motif tables 1and 2. The following format is used for representation of each sequence motif table. Instead of using existing format in this paper protein logo representation has been used [3]. The top box shows the number of sequence segments belonging to this motif, percentage of structural similarity, and average HSSP-BLOSUm62 value. The graph demonstrates the type of amino acid frequently appearing in the given position by amino acid logo.

Motif Table 1	Motif Table 2
Helices Motif with conserved A, E	Helices Motif with conserved K,E
Number of Sequence Segments:1364	Number of Sequence Segments: 1241
Structural Similarity: 81.3123	Structural Similarity: 76.7284
HSSP-BLOSUM62: 0.8234	HSSP-BLOSUM62: 0.6043
	

6 Conclusion

The K-Means algorithm is commonly used for large datasets. But, the benchmark algorithm do not always guarantee good results as the accuracy of final clusters depends on the selection of initial centriods. In this paper, an attempt has been made to select seeds for K-Means Clustering algorithm. The proposed seeding method helps us to identify more number of sequence motifs that are hidden inside proteins in an efficient manner. Cluster compactness of our proposed seeding technique outputs is satisfactory as well. Our future work aims by extending this frame work with methods for refining the computation of initial centroids.

References

[1] Attwood, T.K., Beck, M.E., Bleasby, A.J., Degtyarenko, K., Smith, D.J.P.: Progress with the PRINTS protein fingerprint database. Nucleic Acids Res. 24, 182–183 (1996)
[2] Alter, O., Brown, P.O., Boststein, D.: Singular value decomposition for genome-wide expression data preprocessing and modelling. PNAS 97(18), 10101–10106 (2000)
[3] Chen, B., Tai, P.C., Harrison, R., Pan, Y.: FGK Model: An Efficient Granular Computing Model for Protein Sequence Motifs Information Discovery. In: IASTED Proc. International Conference on Computational and Systems Biology(CASB), Dallas, pp. 56–61 (2006)
[4] Cox, E.: Fuzzy Modelling and Genetic Algorithms for Data Mining Exploration. Elsevier (2005)
[5] Henikoff, S., Henikoff, J.G., Pietrokovski, S.: Blocks+: a non redundant data-base of protein Alignment blocks derived from multiple compilation. Bioinformatics 15(6), 417–479 (1999)
[6] Hullo, N., Sigrist, C.J.A., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A.: Recent improvements to the PROSITE database. Nucleic Acids Res. 32(database issue), D134–D137 (2004)
[7] Kabsch, W., Sander, C.: Dictionary of protein secondary structure pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983)
[8] Sander, C., Schneider, R.: Database of Homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Struct. Funct. Genet. 9(1), 56–68 (1991)
[9] Wang, G., Dunbrack Jr., R.L.: PISCES: a protein sequence culling server. Bioinformatics 19(12), 1589–1591 (2003)
[10] Zhong, W., Altun, G., Harrison, R., Tai, P.C., Pan, Y.: Improved K-Means Clustering algorithm for Exploring Local Protein Sequence motifs Representing Common Structural Property. IEEE Transactions on Nanobioscience 4(3), 255–265 (2000)

A Proposed Hybrid Medoid Shift with K-Means (HMSK) Segmentation Algorithm to Detect Tumor and Organs for Effective Radiotherapy

V.V. Gomathi[1] and S. Karthikeyan[2,*]

[1] Research and Development Centre, Bharathiar University, Coimbatore, India
vv.gomathi@gmail.com
[2] Department of Information Technology, College of Applied Sciences, Sohar, Oman
skaarthi@gmail.com

Abstract. Image segmentation plays a significant role in many medical imaging applications. Manual segmentation of medical image by the radiologist is not only a tiresome and time consuming process, also not a very accurate with the increasing medical imaging modalities and unmanageable quantity of medical images that need to be examined. Therefore it is essential to examine current methodologies of image segmentation. Enormous research has been done in creating many different approaches and algorithms for medical image segmentation, but it is still difficult to evaluate all the images. However the problem remains challenging, with no general and unique solution. This paper reviews some existing medical image segmentation algorithms suitable for CT images. Their pros and cons were analyzed and proposed a HMSK algorithm for slices of CT images to give effective radiation therapy.

Keywords: Computed Tomography, Segmentation, Active Contour Model, Watershed Segmentation, K-Means, Fuzzy CMeans, Mean Shift Segmentation, Medoid Shift Segmentation.

1 Introduction

Image segmentation is the process of partitioning a digital image into segments. Segmentation refers to simplifying and/or change the representation of an image into more meaningful and easier to analyze [8]. Image Segmentation is the most interesting and challenging problems in computer vision generally and especially in medical imaging applications. Accurate, fast and reproducible image segmentation techniques are required for various applications. Segmentation algorithms available vary widely depending on the specific application, image modality and other factors. Medical image segmentation is the process of outlining relevant anatomical structures in an image dataset. It is a problem that is central to a variety of medical applications including image enhancement and reconstruction, surgical planning, disease classification, data storage and compression, and 3D visualization [12].

* Corresponding author.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 139–147, 2013.
© Springer International Publishing Switzerland 2013

The segmentation of the medical image faces the challenges of the existence in a large number of diverse structures of human anatomy and inevitable artifacts induced from the imaging procedure. Most medical image segmentation algorithms are either semi-automatic or require some form of human intervention to perform satisfactorily.

One fundamental problem in medical image analysis is image segmentation which identifies the edges of objects such as organs or abnormal regions (e.g. tumors) in images. Good segmentations will help clinicians and patients as they provide vital information for 3-D visualization, surgical planning and early disease recognition. The most common images used in medicine, particularly in radiation therapy are the Computer Tomography Images (CT), because of its electron density. It is a diagnostic tool used to detect cancer and find out the cancer's stage. It has the gold standard for radiation therapy planning. CT has the most well defined source detector geometry and also has the highest contrast rate than other modalities.

The Second leading causes of death in the world are different types of cancer. According to World Health Organization Statistics cancer is responsible for 7.6 million deaths worldwide annually. Deaths from cancer worldwide are projected to continue rising, with an estimated 13.1 million deaths in 2030. The most useful way to reduce deaths due to cancer is to treat the disease in the early stages. Early treatment needs early diagnosis, and early diagnosis requires an accurate and reliable diagnostic procedure that allows physicians to differentiate tumors from normal organs. Manual segmentation process requires at least three hours completing by the radiologist or radiation oncologist. It is time-consuming process and also inter- and intra-expert variability residing between persons to persons. Hence, manual segmentation becomes less efficient in clinical operations and human interpretation may not be produce correct information. So intellectual techniques are needed to segment automatically.

The problem addressed in this paper is how to automate the organ segmentation in three dimensions from CT volume images. In this paper different medical image segmentation algorithm such as Active Contour Model, Watershed Transform, K-Means Clustering, Texture Based Algorithm, Mean Shift Algorithm and Medoid Shift Algorithm were analyzed and propose a HMSK algorithm. The Experimental result proves that the proposed method gives promising results.

2 Related Works

Kaihua Zhang et al proposed a novel region-based active contour model which is implemented with special processing named Selective Binary and Gaussian Filtering Regularized Level Set (SBGFRLS) method [6]. LiWang et al proposed an improved region-based active contour model in a variational level set formulation. They defined energy functional with a local intensity fitting term, which is dominant near object boundaries and responsible for attracting the contour toward object boundaries, and an auxiliary global intensity fitting term, which incorporates global image information to improve the robustness of the proposed method. Our model can handle intensity homogeneity, and allows for flexible initialization [9]. Yang Xiang et al proposed a new edge-based active contour method, which uses a long range and orientation-dependent interaction between image boundaries and the moving curves while maintaining the edge fidelity. This method is able to detect sharp features of the

images [15]. Wenbing Tao et al proposed an improved variational model, multiple piecewise constant with geodesic active contour (MPC-GAC) model, which generalizes the region-based active contour model and merges the edge-based active contour to inherit the advantages of region-based and edge-based image segmentation models [14]. Jianhua Liu et al proposed watershed algorithm and region merging approach, to solve over-segmentation for identifying Liver Cancer CT image [5]. Hassan Masoumia et al proposed a combined algorithm that utilizes MLP neural networks and watershed algorithm to avoid over segmentation. The proposed algorithm extracts liver region in one slice of the MRI images [4]. Shojaii et al presented a novel lung segmentation technique based on watershed transform which eliminates the tasks of finding an optimal threshold and separating the attached left and right lungs, which are two common practices in most lung segmentation methods and require a significant amount of time [11]. Yu-Len Huang et al have used neural network classification and morphological watershed segmentation to extract precise contours of breast tumors from US images [16].

Elmasry et al presented and evaluated the performance of K-means and normalized cuts segmentation approaches that were applied to liver computed tomography (CT) images. They have observed that K-means clustering algorithm outperformed normalized cut segmentation algorithm for cases where region of interest depicts a closed shape, while, normalized cut algorithm obtained better results with non-circular clusters [3]. Mounir Sayadi et al proposed a cascade clustering method combining statistical features and the standard Fuzzy C-Means clustering algorithm. Instead of using the gray level value of a given pixel, a feature vector is extracted from a sliding window centered on the pixel. The Fuzzy C-means algorithm is used to cluster the obtained feature vectors into several classes corresponding to the different regions of the multi-textured image [10]. Ehsan Nadernejad et al have presented a Fuzzy C-Means algorithm for segmenting Pixonal images [2]. Keh-Shih Chuang et al presented a fuzzy c-means (FCM) algorithm that incorporates spatial information into the membership function for clustering. This method yields regions more homogeneous than those of other methods and also reduces the spurious blobs and removes noisy spots, this technique is a powerful method for noisy image segmentation and works for both single and multiple-feature data with spatial information [7].

Ali Hassan Al-Fayadh et al proposed a Mean Shift Algorithm and active contour to detect objects for CT Angiography Image Segmentation. This method of boundary detection together with the mean-shift can achieve fast and robust tracking of the CT Angiography Image Segmentation in noisy environments [1]. Wenbing Tao et al developed a novel approach that provides effective and robust segmentation of color images by incorporating the advantages of the mean shift (MS) segmentation and the normalized cut (Ncut) partitioning methods [13]. Arnaldo Mayer et al presented an automated segmentation framework for brain MRI volumes based on adaptive mean-shift clustering in the joint spatial and intensity feature space [8].

3 Limitation of the Existing Algorithm

Each segmentation algorithm has its own limitations, where any of these algorithm shall be developed to get accurate segmentation, but comparing the algorithms for the

application of tumor and organs detection , relying on these algorithms shall not be the opt solution. The Region based watershed segmentation method involves in over segmentation and when controlled through clustering and morphology operation, repeatability could not be achieved. In case of Active contour method is performed then the intensity difference of organs varies from each other, where developing further in this process shall involve lots of work and accurate results shall not be predicted. In the case of K-Means algorithm, the noise could be segmented as an organ. FCM shall only generate good result for certain organ which have notable differences in gray scale value.

In case of mean shift segmentation the parameter value involving in algorithm cannot be fixed for all the organs. In these cases, medoid shift clustering algorithms have found to be suitable. But the image pixel value would get clustered in its limited area only and there were additional noise influencing the segmented components (over fragmentation), of the particular connected component could not be acquired.

Medoid segmentation shall lead to over fragmentation. This occurs because of clustering the image data for very low gray scale value difference. Heart and liver regions differ very minimal in gray value, every small difference needs to be clustered, which result in over fragmentation (lot of connected components in the required region). The proposed HMSK algorithm did overcome the over fragmentation issues.

4 Proposed Methodology

An efficient HMSK algorithm is proposed to avoid the drawbacks of medoid shift algorithm. One of the major drawbacks was over fragmentation produced by the medoid shift algorithm. This will produce different connected components. So classifying many connected components is not possible. This will cause wrong segmentation and also lead to wrong decision making by the radiologist. Here HMSK is designed to avoid the above said limitation and segmenting the required organs efficiently.

HMSK Algorithm

The Proposed HMSK Algorithm is as follows:

Step I: Consider the Single Dicom image or slices of Dicom images

Step II: Apply the ECFT (Enhanced Curvelet Filter Technique) algorithm to get a noiseless image

Step III: Find the Histogram of the input image

Step IV: Initialize the control parameter

Step V: Find the gray level cluster values based on an initialized control parameter

Step V: Find no of pixel values present between each range of all gray level cluster values.

Step VI: Cluster the pixels which lies between the range to the respective gray level cluster value

Step VII: Each cluster are considered as data points

Step VIII: Find the distance between each cluster to all the data points

Step IX: Make the data point allocation by using $\left(\sum_{(i=1)}^{c} * \sum_{(j=1)}^{c} \left(\left\|\left(x_i - v_j\right)\right\|\right)^2\right)$

Step X: Find the new cluster center by using $\left(1/C_i\right) * \left(\sum_{j=1}^{C_i} \left(x_i\right)\right)$

Step XI: Obtain the Clustered Image.

5 Data Set

Different type of Tumor patient dataset was collected by a SIEMENS SOMATOM EMOTION SPIRAL CT scanner located at Multi Speciality Hospital, Coimbatore. Besides a normal scan performed at a routine clinical dosage (130 mA), an additional scan from the same patient was acquired at a much lower tube current, i.e. 20 mA.

6 Experimental Results and Discussion

Experimentation was carried out on 100 numbers of different tumor patients contains 100 to 1000 slices of Computer Tomographic images using different Segmentation algorithms. The image format is DICOM (Digital Imaging Communications in Medicine). The proposed algorithm has been implemented in Matlab environment. Manual Segmentation done by the experts. Experimental results of the images are illustrated here.

Patient ID 002

Fig. 1. Output of Manual Segmentation

Patient ID 002:
The results produced by Various Segmentation Algorithms:

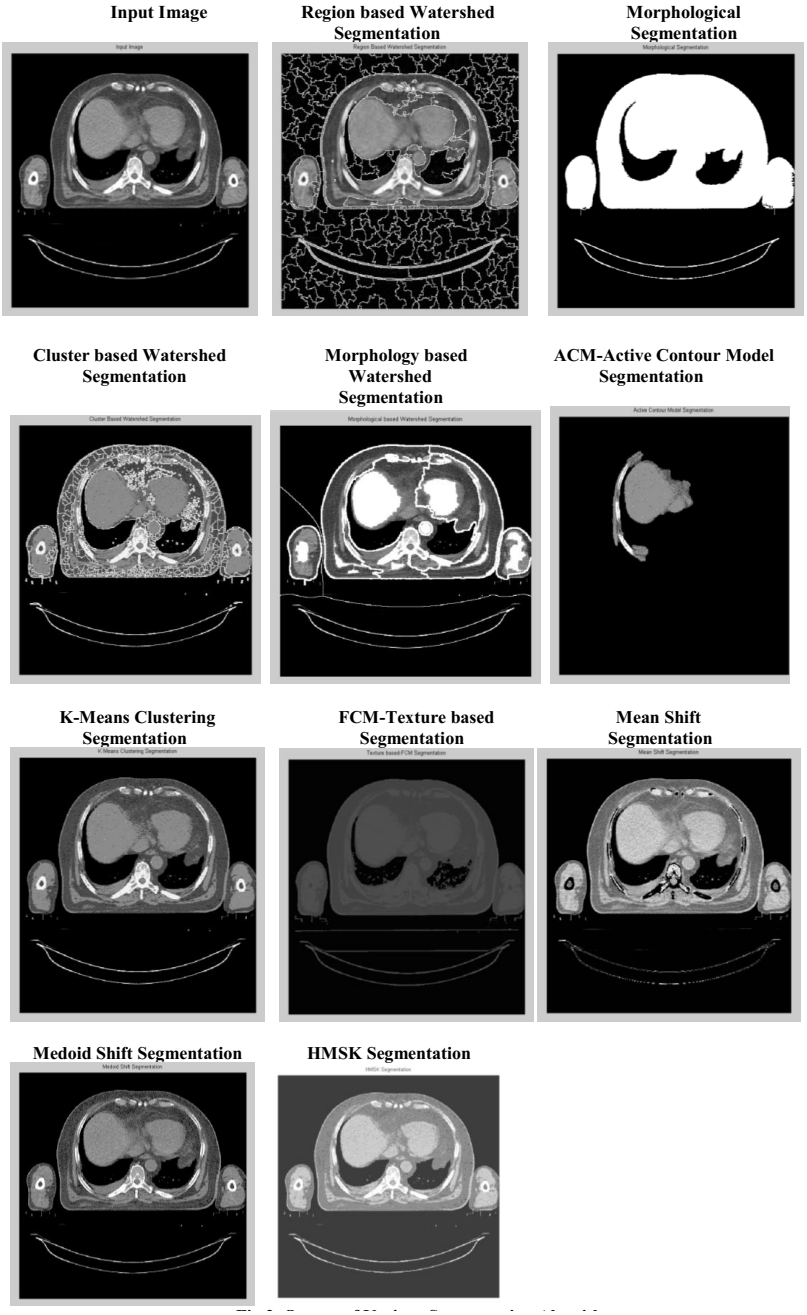


Fig. 2. Output of Various Segmentation Algorithms

6.1 Discussion

The main goal of the proposed research is to overcome the limitations of the existing algorithms mainly decrease the connected Components and reducing the over fragmentation in the image segmentation. The proposed HMSK performs well and furnish significant results. The above output figures depicts the result of HMSK is very clear and accurate.

Segmentation, involves the output with over segmented data, because the variation in the image is more. When the clustered image is given as input, then the variation of image could not be able to detect properly because there are lots of image value getting changed. The parameter values could not be able to fix for the algorithm. In case of morphology based, again the noise interrupts the entire process. Inactive contour segmentation the same happens, the algorithm parameter cannot be adjusted for all areas to mark the component edges depending upon the image variations.

The algorithms shall be developed for particular intensity of images and at the outset the contrast enhancement algorithm is performed. This algorithm initially maintains the contrast level common for all the input images. Maintaining the contrast level will be the same but some of the patient image results in drastic changes of values beyond the normal. It would be difficult and the result of acquiring the values for algorithm parameters is remote.

In case of clustering based algorithm the effect of noise would get neutralized. The problem arrived at clustering based segmentation is grouping of two connected components into one. K-means clustering have to be with the input of initial clusters, FCM generates the output for one particular set of images, determining a segmentation of all organs were remote, in case of mean shift clustering the algorithm values could not be fixed in common for effective segmentation. Medoid shift algorithm suitable among the entire clustering algorithm. This particular algorithm is modified by the factor called control parameter. The control parameter holds the minimum difference between the organ and the background. However the output of the same induced noises gives the appearance of a component. But the results produced by the proposed HMSK are most significant. The comparison result is shown in Table1.

Table 1. Quantitative Analysis for all Segmentation Methods

S.No	Segmentation Methods	Quantitative Parameters			
		Sensitivity(%)	Specificity(%)	Accuracy(%)	Fragments
1	Region based Watershed Segmentation	99.40	96.70	94.79	4571
2	Cluster based Watershed	99.36	99.93	95.02	4571
3	Morphological Watershed	99.91	96.64	94.76	3578
4	ACM	98.81	99.14	97.12	511
5	K-Means	99.79	96.85	94.95	3578
6	FCM	97.96	96.96	94.99	2172
7	Mean Shift	98.12	97.96	96.03	11400
8	Medoid Shift	99.71	97.23	95.31	8979
9	**Proposed HMSK**	**97.41**	**99.41**	**99.34**	**349**

Evaluating the results produced by segmentation algorithms is challenging task. The segmentation is evaluated by assessing its consistency with the manual segmentation and their amounts of fragmentation. The value of Number of fragments indicates the number of connected components in the required region to identify as organ. Ultimately the value should be low for best segmentation method, since more connected components shall lead to inaccurate organ recognition and classification. In this research Table 1 depicts the sensitivity, specificity, accuracy and numbers of Fragments are considered to compare the performance of various segmentation algorithms. Among the entire segmentation algorithm HMSK produces higher accuracy and lower the number of fragments. It is clearly shows that segmentation results of proposed HMSK are most promising.

7 Conclusion

In this paper we have analyzed various segmentation algorithms which are most suitable for CT image segmentation especially for tumor. Cluster based segmentation algorithms are suitable for CT image segmentation. The Complexity of the medoid shift algorithm is lower than mean shift. But still over fragmentation exists in medoid shift Segmentation algorithm. This over fragmentation gives many clusters and it consider as organ. This will lead to give wrong therapy. Therefore we proposed a HMSK algorithm to avoid over fragmentation. This algorithm is segmenting the organ accurately. It is also very competitive; resulting in good segmentations compared with the existing segmentation algorithms.

Acknowledgement. The authors like to thank Dr.M.Hemalatha, Professor, Department of Computer Science, Karpagam University for her valuable suggestions, comments and words of encouragement. It helped us to make this research work successful.

References

1. Al-Fayadh, A.H., Mohamed, H.R., Al-Shimsah, R.S.: CT Angiography Image Segmentation by Mean Shift Algorithm and Contour with Connected Components Image. International Journal of Scientific & Engineering Research 3(8) (2012) ISSN 2229-5518
2. Nadernejad, E., Sharifzadeh, S.: A New method for image segmentation based on Fuzzy C-Means algorithm on pixonal images formed by bilateral filtering. Springer (2011)
3. Moftah, E., El-Bendary, Hassanien: Performance evaluation of computed tomography liver image segmentation approaches. In: International Conference on Hybrid Intelligent Systems, HIS (December 2012)
4. Masoumia, H., Behradb, A., Pourminaa, M.A., Roostac, A.: Automatic liver segmentation in MRI images using an iterative watershed algorithm and artificial neural network. Biomedical Signal Processing and Control (2012)
5. Liu, J., Wang, Z., Zhang, R.: Liver Cancer CT Image Segmentation Methods Based on Watershed Algorithm. In: International Conference on Computational Intelligence and Software Engineering (December 2009)

6. Zhang, K., Zhang, L., Song, H., Zhou, W.: Active contours with selective local or global segmentation: A new formulation and level set method. Image and Vision Computing, 668–676 (2010)
7. Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J., Chen, T.-J.: Fuzzy c-means clustering with spatial information for image segmentation. Computerized Medical Imaging and Graphics, 9–15 (2006)
8. Shapiro, L.G., Stockman, G.C.: Computer Vision, pp. 279–325. Prentice-Hall, New Jersey (2001) ISBN 0-13-030796-3
9. Wang, L., Li, C., Sun, Q., Xia, D., Kao, C.-Y.: Active contours driven by local and global intensity fitting energy with application to brain MR image segmentation. Computerized Medical Imaging and Graphics, 520–531 (2009)
10. Sayadi, M., Tlig, L., Fnaiech, F.: A New Texture Segmentation Method Based on the Fuzzy C-Mean Algorithm and Statistical Features. Applied Mathematical Sciences 1 (2007)
11. Shojaii, R., Alirezaie, J., Babyn, P.: Automatic lung segmentation in CT images using watershed transform. In: International Conference on Image Processing (2005)
12. Tsai, A., Wells, W., Tempany, C., Grimson, E., Willsky, A.: Mutual information in coupled multi-shape model for medical image segmentation. Medical Image Analysis, 429–445 (2004)
13. Tao, W., Jin, H., Zhang, Y.: Color Image Segmentation Based on Mean Shift and Normalized Cuts. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics 37(5) (2007)
14. Tao, W., Tai, X.-C.: Multiple piecewise constant with geodesic active contours (MPC-GAC) framework for interactive image segmentation using graph cut optimization. Image and Vision Computing, 499–508 (2011)
15. Xiang, Y., Chung, A.C.S., Ye, J.: An active contour model for image segmentation based on elastic interaction. Journal of Computational Physics, 455–476 (2006)
16. Huang, Y.-L., Chen, D.-R.: Watershed Segmentation For Breast Tumor In 2-D Sonography. Ultrasound in Med. & Biol. 30(5), 625–632 (2000)

Face Representation Using Averaged Wavelet, Micro Patterns and Recognition Using RBF Network

Thangairulappan Kathirvalavakumar[1] and J. Jebakumari Beulah Vasanthi[2]

[1] Department of Computer Science, V.H.N.S.N College,
Virudhunagar, 626001, TN, India
Kathirvalavakumar@yahoo.com
[2] Department of Computer Applications, A.N.J.A College,
Sivakasi, TN, India
jebaarul07@yahoo.com

Abstract. Recognition of human faces is a very important task in many applications such as authentication and surveillance. An efficient face recognition system with face image representation using averaged wavelet and wavelet packet coefficients, Discriminative Common Vector (DCV) and modified Local Binary Patterns (LBP) and recognition using radial basis function (RBF) network is presented. Face images are decomposed by 2-level wavelet and wavelet packet transformation. The discriminative common vectors are obtained for averaged wavelet. The new proposed LBP operator is applied on the obtained DCV and also applied on averaged wavelet packet coefficients of all the samples of a class. The histogram values obtained from the LBP are recognized using RBF network. The proposed work is tested on three face databases such as Olivetti Oracle Research Lab (ORL), Japanese Female Facial Expression (JAFFE) and Essex face database. The proposed method results in good recognition rates along with less training time because of the extracted discriminant input from the preprocessing steps involved in the proposed work.

Keywords: Face recognition, Wavelet, Wavelet packets, Discriminative common vector Classification, Local binary patterns, Radial basis function network.

1 Introduction

Recently, due to military, commercial and law enforcement applications there has been much interest in automatically recognizing faces in still and video images. The data come from wide variety of sources mainly passports, credit cards, photo ID, driver's licenses and mud shots. A variety of changes in face images also a great challenge. A face recognition system must be robust with respect to the much variability of face images such as viewpoint, illumination, and facial expression. The two main tasks in the face recognition system are representation of face and classification of the face.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 148–159, 2013.
© Springer International Publishing Switzerland 2013

Many methods have been proposed for face recognition in which appearance based approaches operates directly on images or appearances of face objects. Turk and Pentland have implemented the Eigenfaces approach [15] which is the popular face recognition method. The approach transforms face images into a small set of characteristic feature images, called eigenfaces, which are the principal components of face images. Linear Discriminant Analysis (LDA) could be operated either on the raw face image to extract the Fisherface [2] or on the eigenface to obtain the discriminant eigen features [14]. The discriminative common vectors algorithm is a recently addressed discriminant method, which shows better face recognition effects than some commonly used linear discriminant algorithms [17].

The wavelet techniques are applied to solve many real world problems, particularly in image processing and face recognition [7, 11]. An appropriate wavelet transform can result in robust representations with regard to lighting changes and be capable of capturing substantial facial features while keeping computational complexity low [20]. Wavelet Packet Decomposition (WPD) of an image is a useful technique for building compact and meaningful feature vector and is also used in face recognition [8]. Perlibakas [16] has presented a work on face recognition using both principal component analysis and wavelet packet decomposition. Wong *et al*. [18] have proposed wavelet packet transformation to decompose images into frequency subbands.

Cevikalp *et al*. [4] have proposed a new face recognition method called the discriminative common vector, in which one algorithm uses within-class scatter matrix while the other uses the subspace method and the Gram-Schmidt orthogonalization procedure to obtain the discriminative common vectors. Cevikalp et al. have proposed a DCV method with kernels, which first map all data samples to a higher dimensional feature space through a nonlinear mapping, and then,the DCV method is applied in the mapped space [5]. Carlos et al. have performed the feature reduction based on discriminative common vector which result in less load time and improved recognition rate [3].

LBP features were proposed originally for texture analysis [12] and recently have been introduced to represent faces in face recognition and face detection The idea of using local binary patterns for face description is motivated by the fact that faces can be seen as a composition of micro patterns which are properly described by this operator and, it has become a very popular technique in recent years. Ahonen [1] have presented a novel approach based on local binary pattern histograms for face recognition, which considered both shape and texture information in representing face images. Pujol and Gracia have presented a method to calculate the most important principal LBP for recognizing faces [13]. Zhou *et al*. [21] have proposed a face representation approach which fuses Gabor filter, Local Binary Pattern and Local Phase Quantization. Radial Basis Function neural networks (RBF) are suitable for pattern recognition and classification and used in face recognition [6]. Li and Yin have presented a face recognition system using radial basis function neural network and wavelet transformation [11]. A radial basis function neural network with a new incremental learning method

based on the regularized orthogonal least square (ROLS) algorithm is proposed by Wong [19] for face recognition. A face recognition approach based on kernel discriminative common vectors (KDCV) and RBF network is proposed [9].

In this paper, an efficient face recognition system with face image representation using averaged wavelet and wavelet packet coefficients, discriminative common vector and modified local binary patterns and recognition using radial basis function neural network is presented. The rest of the paper is structured as follows: the next section describes subsequent steps in the face representation stage. Section 3 presents the proposed recognition process using radial basis function network. Section 4 describes the data set and experiment results along with discussions.

2 Face Representation

This proposed work comprises of an effective face representation scheme using wavelet, wavelet packet transformation, DCV and LBP operator.

2.1 Wavelet Transformation

Wavelet transformation results in a strong representation with regard to lighting changes and capable of capturing substantial facial features. In the proposed work, wavelet transform is created by passing the face image through a series of filter bank stages. In level-1 of wavelet transformation, the face image is applied to low pass and high pass filter and the values are downsampled by a factor in the horizontal direction. In level-2, the filtered output of level-1 is then filtered by an identical filter pair in the vertical direction. The decomposition of the image into four frequency subbands is denoted by approximation (LL) and detailed coefficients (HL, LH, and HH). In this proposed work,the low frequency approximation wavelet coefficients are only considered for further processing.

2.2 Wavelet Packet Transformation

In the proposed work, the wavelet packet transformation is also used. In wavelet transformation, a face image is decomposed into approximation and detailed coefficients and then the approximation coefficient is only used for the further level decomposition. Unlike wavelet transformation, in the wavelet packet transformation, approximation coefficients as well as detailed coefficients of the face are used in the second level for decomposition and so on. The wavelet packet transformation is illustrated using the wavelet packet tree as in Fig. 1[16].

The original image A_0^0 of level 0 is decomposed into approximation A_0^1, horizontal details D_{0h}^1, vertical details D_{0v}^1 and diagonal details D_{0d}^1 at level 1. The approximation A_0^1, is further decomposed into the nodes such as A_0^2, D_{0h}^2, D_{0v}^2, D_{0d}^2, in the level 2. Likewise, horizontal, vertical and diagonal details are decomposed into the nodes such as A_1^2, D_{1h}^2, D_{1v}^2, D_{1d}^2, A_2^2, D_{2h}^2, D_{2v}^2, D_{2d}^2, A_3^2, D_{3h}^2, D_{3v}^2 and D_{3d}^2 respectively.

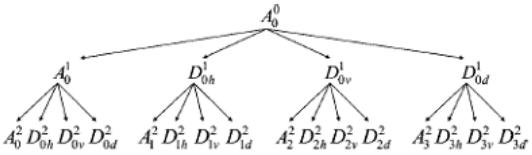


Fig. 1. Wavelet packet tree

In order to get decomposition at level l, decompose the approximations A_i^{l-1}, $D_{i,h}^{l-1}$, $D_{i,d}^{l-1}$ and $D_{i,v}^{l-1}$ as follows as proposed by Perlibakas [16]:

$A_i^{l-1} \rightarrow \{A_{4i}^l; D_{4i,h}^l; D_{4i,v}^l; D_{4i,d}^l\}, l > 0$

$D_{i,h}^{l-1} \rightarrow \{A_{4i+1}^l; D_{4i+1,h}^l; D_{4i+1,v}^l; D_{4i+1,d}^l\}, l > 0,$

$D_{i,v}^{l-1} \rightarrow \{A_{4i+2}^l; D_{4i+2,h}^l; D_{4i+2,v}^l; D_{4i+2,d}^l\}, l > 0,$

$D_{i,d}^{l-1} \rightarrow \{A_{4i+3}^l; D_{4i+3,h}^l; D_{4i+3,v}^l; D_{4i+3,d}^l\}, l > 0,$

where $i = 0, ..., (4^{l-1} - 1)$ During the wavelet transformation, at level 2, only the approximation is decomposed into approximation and details.The wavelet packet transformation has been applied on the face images for two levels and obtains four wavelet packet coefficients namely approximation, horizontal details, vertical details, and diagonal details. The obtained wavelet packets are processed using two different methods and the resultant is used as an input for recognition.

2.3 Discriminative Common Vectors

Wavelet coefficients are applied directly in most of the applications. The direct use of wavelet coefficients may not extract most discriminative features for two reasons such as much redundant or irrelevant information contained in wavelet coefficients and also unable to acquire new meaning underlying features which have more discriminative power. In order to overcome the deficiency of direct use of wavelet coefficients, it is proposed to construct discriminant features from the wavelet coefficients using within-class scatter matrix method. A common vector for each individual class is obtained by removing all the features that are in the direction of the eigenvectors corresponding to the nonzero eigen values of within-class scatter matrix of all classes. The new set of vectors, called the discriminative common vectors, are used for recognition.

Let the training set be composed of C classes, where each class contains N sample, and let x^{i_m} denotes the m^{th} sample from the i^{th} class. Within-class scatter matrix of the sample is constructed to obtain the feature vectors which is defined as [4].

$$S_w = BB^T \tag{1}$$

where the matrix B is given by

$$B = [x_1^1 - \mu_1, ..., x_N^1 - \mu_1, x_1^2 - \mu_2, ..., x_N^C - \mu_C] \tag{2}$$

where x_i^j is i^{th} sample of class j and μ_j is mean of the samples in the j^{th} class.

Let $Q = [\alpha 1, ...\alpha r]$, is the set of orthonormal eigenvectors corresponding to the non-null eigenvalues of S_w and r is the dimension of S_w. Next project the input sample on the null space of S_w in order to get the common vectors x_{com} defined as:

$$x_{com}^i = x_m^i - Q\overline{Q}x_m^i \tag{3}$$

where $m = 1...N$samples and $i = 1...C$ classes. Calculate the principal components of S_{com} (the eigenvectors w_k), which correspond to the non zero eigenvalues, by

$$J(W_{opt}) = \arg\max w[W^T S_{com} W] \tag{4}$$

where S_{com} is computed as

$$S_{com} = B_{com} B_{com}^T \tag{5}$$

where B_{com} is given by

$$B_{com} = [x_{com}^1 - \mu_{com}, ..., x_{com}^C - \mu_{com}] \tag{6}$$

The Feature Vector of Training set is calculated as

$$\Omega_i = W^T x_m^i \tag{7}$$

Similarly, to recognize a test image x_{test}, the feature vector of the test image is found by

$$\Omega_{test} = W^T x^{test} \tag{8}$$

The procedure for finding the discriminative common vector is summarized below.

1. Compute nonzero eigenvalues along the corresponding eigenvectors of S_w using the resultant matrix of BB^T, where B is computed using equation (2).
2. Choose an input sample from each class and project it onto the null space of S_w to obtain the common vectors. Compute x_{com}^i using equation (3).
3. Compute eigenvectors w_k of S_{com}, corresponding to the nonzero eigenvalues, using the equations (4) and (5).
4. The feature vector for training set and test set is obtained respectively from equation (7) and (8).

2.4 Local Binary Pattern

The LBP operator was introduced by Ojala [12]. The idea of using the LBP features is that the face image can be seen as a composition of micro-patterns. The LBP operator defined in 3x3 neighbourhood is shown in Fig. 2. "11010011" is the designed pattern of the central pixel. In this proposed work, a modified computation proposed by Kathirvalavakumar and Vasanthi [10] is used for finding the LBP. For each pixel except the edge pixels, the binary bitmap is found using 3x3 neighborhood elements. The average value for 3x3 map is computed. The averaged value is used as a threshold and compared with each element of

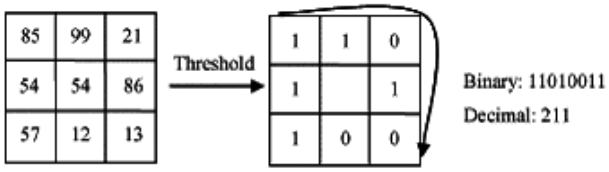


Fig. 2. Local binary pattern operator

the 3x3 map including the central point. Binary map 1 value is set as 1 if the element is greater than the averaged value, otherwise 0. The new LBP operator is illustrated in Fig. 3. A binary pattern is obtained by concatenating all these binary values by a row major starts from left. The corresponding decimal value of the generated binary number is then used for labeling the given pixel. The obtained binary numbers are referred as the LBPs. In Fig. 3, the local binary pattern is 001011010 and its decimal equivalent is 90. The histogram values are to be generated for recognizing the resultant face representation of LBP operation. The histogram values are supplied as input pattern to the RBF for recognition. Proposed work includes two methods in the preprocessing which

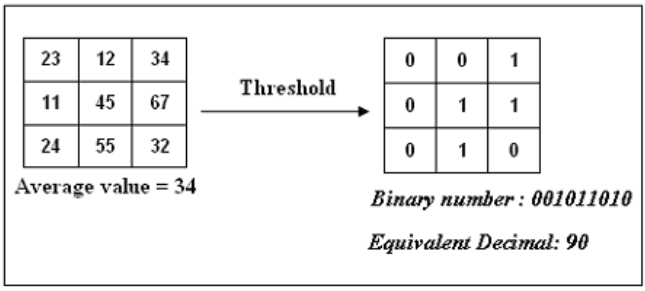


Fig. 3. New Local binary pattern operator defined in 3 x 3 neighbourhood

use wavelet and wavelet packet transformations. In the Method - I, the face images are decomposed by 2-level wavelet transformation. The resultant adjacent wavelet coefficients are averaged rowwise. The discriminative common vectors are obtained for averaged wavelet coefficients. LBPs are obtained using the new proposed LBP operator and then histogram values are generated. In the Method - II, wavelet packet transformation has been applied on the input face images. The averaged approximation wavelet packet coefficients of all the samples of a class are averaged. DCV and LBP values are computed for the averaged wavelet packet coefficients as similar to the first method. The steps in the Method - I are

1. Perform steps 2 - 4 for all the training samples.
2. Compute wavelet Coefficients by applying wavelet transformation. Average the adjacent wavelet coefficients in rowwise.

3. Compute DCV Coefficients for the resultant averaged wavelet coefficients using the within-class scatter matrix method.
4. Find the Local Binary Pattern using the proposed computation and the histogram values.
5. Train the RBF network for recognizing the values obtained from step 4 using algorithm in section 3.1.
6. Repeat the steps 2-5 for the test set face images.
7. Classify the DCV coefficients of test image using trained RBF network.

The steps in the Method - II are listed below.

1. Apply 2-level wavelet packet transformation on a sample of a class of training set and compute the wavelet packet coefficients.
2. Add the four child nodes of A_0^1 and compute its average.
3. Repeat steps 1 and 2 for all the training samples of the class.
4. Find the average of averaged approximation wavelet coefficients of the class.
5. Repeat steps 1 to 4 for all classes.
6. Find the Local Binary Pattern using the proposed computation and the histogram values.
7. Train the RBF network for recognizing the values obtained from step6 using algorithm in section 3.1.

3 Recognition by Radial Basis Function Network

In general,the RBF network contains three layers: input, hidden and output layers. The input neurons are normalized and then feed the values to each of the neurons in the hidden layer. The units of the hidden layer usually correspond to the clustering centers of the training sample set. The basis function of the hidden layer neurons are considered to be Gaussian function and computed basis function output are passed to the output layer. The hidden layer output is computed as

$$\phi_j(X) = \exp\{\frac{||X-\mu||^2}{\sigma^2}\} \tag{9}$$

where $X = [x_1, x_2, ..., x_n]^T$ is the input vector, μ is the center, and σ is the width. The output layer represents the outputs of the network, and each output node is a linear combination of the k radial basis functions of hidden nodes. The output of the RBF is computed as

$$y_i = \sum_{j=1}^{k} W_{ji}\phi_j(X) \tag{10}$$

where W_{ji} are the weights connecting the hidden layer neuron j and output layer neuron i. The weights are adjusted according to the formula,

$$w_{t+1} = w_t + \lambda(d_i - y_i)\phi_j(X) \tag{11}$$

where λ is a positive learning rate parameter and d_i is the desired output.

3.1 RBF Training Algorithm

The training algorithm of Radial Basis Function Network is given as follows.

1. Generate random number to initialize the weights of the RBF network.
2. For a input pattern compute hidden layer output using the equation (9).
3. Compute the output layer output using the equation (10).
4. Find the error as the difference between desired and actual output obtained.
5. Adjust the hidden layer weights according to equation (11)
6. Repeat steps 2-5 for all input patterns.
7. Compute sum of squared error of the network.
8. Repeat steps 2-7 until the acceptable minimum error level is reached.

4 Results and Discussions

The proposed system is tested using face databases such as ORL, The Japanese Female Facial Expression (JAFFE) and Essex Face database. The ORL face data base contains 40 faces of size as 112 x 92 and each face has 10 different facial views representing various expressions, small occlusion by glasses, different scale and orientations. Hence, there are 400 face images in the database. In the proposed methods, 5 different poses of 20 person's faces are used for training and 5 different poses of 20 person's faces are used for testing. The Japanese Female Facial Expression contains different facial expressions posed by Japanese female models in which 5 different poses of 15 models are used for training and the other 5 different poses of 15 models are used for testing. The actual dimension of the image is 256 x 256. The Essex Face database is having faces of more than 150 male and female with 20 images per individual of University of Essex, UK with the size of 180 x 200. The 5 different poses of 20 person's face are used for the training and 5 different poses of 20 person's faces are used for the testing in the proposed methods. The 5 poses for training and 5 poses for testing are considered sequentially. After applying the wavelet transformation to 2-level using different types of wavelet on the input images of the three databases, the reduced size of the resultant wavelet coefficients is shown in Table 1. The original image, averaged wavelet image and image after applying new LBP operator are shown in Fig. 4. The original samples, averaged wavelet packet image representing the class and image after applying new LBP operator are shown in Fig. 5.

The recognition rates are obtained by doing the training and then testing process repeatedly for 25 times and averaged. Among the several predefined wavelet families, orthogonal wavelets such Haar, Daubechies, Coiflets, and Symlets are used in the proposed work. The recognition rates and training time of the three databases for the wavelets namely Haar, Sym4, Sym8, Db4, Db6, Coif2, Coif4 are shown in Table 2. In the proposed method-I, the highest recognition rate is 98.3% for ORL and 98% for JAFFE and Essex databases when Haar wavelet is used. The lowest rate is 95.1% for ORL, 95.8% for JAFFE and 95.2% for Essex database when Coif4 wavelet is used. The minimum training time is obtained for ORL database when Haar wavelet is used. The recognition rates and training

Table 1. Size of the wavelet coefficients

Wavelet Name	ORL	JAFFE	ESSEX
Haar	28 x 23	64 x 64	45 x 50
Sym4	33 x 28	69 x 69	55 x 50
Sym8	39 x 34	75 x 75	61 x 56
Db4	33 x 28	69 x 69	55 x 50
Db6	36 x 31	72 x 72	58 x 53
Coif2	36 x 31	72 x 72	58 x 53
Coif4	45 x 40	81 x 81	67 x 62

Fig. 4. Original image, averaged Wavelet + new LBP image

time obtained from the proposed method - II for three databases are shown in Table 3. In the ORL database, the recognition rate obtained using method-I for haar wavelet is 98.3% which is greater than the proposed method-II. The training time obtained is less for Haar wavelet in all the three databases.

The recognition rates obtained for the three face databases on applying other methods are compared in Table 4. The recognition rate 97.54% is obtained by the method Push-Pull Marginal Discriminant Analysis on ORL database. The MLA+NM method has the recognition rate of 97% on JAFFE database. When ESSEX database is used, the recognition rate of 97.2% has been achieved by Curvelet with SVM. Eigen face, wavelet face and SOM methods have better results when combined with neural networks (NN) or convolution neural networks. The comparison of training time of two methods when using ORL database is shown in Fig. 6. In the proposed system, two level two dimensional wavelet, wavelet packet transformations are used for dimensionality reduction of the face image. Discriminant values are obtained from the dimensionality reduced image by the DCV method. When the LBP operator [10] is applied on the discriminant values, micro patterns of the face image are obtained. RBF network efficiently recognize the histogram which is obtained from the micro patterns. This good recognition rate is because of the extracted discriminant input from the preprocessing which we have applied.

Fig. 5. Original samples, averaged wavelet packet image of a class + new LBP image

Table 2. Recognition Rates and Training Time of ORL, JAFFE and Essex databases - Method I

Wavelet Name	Recognition Rate (%)			Training Time in Seconds		
	ORL	JAFFE	ESSEX	ORL	JAFFE	ESSEX
Haar	98.3	98.0	98.0	1.856	2.031	1.915
Sym4	97.6	97.1	96.4	2.241	2.326	2.287
Sym8	97.2	97.1	96.6	2.345	2.407	2.379
Db4	97.6	97.2	96.8	2.286	2.388	2.301
Db6	96.0	96.6	96.4	2.439	2.603	2.476
Coif2	96.4	96.2	97.3	2.382	2.485	2.415
Coif4	95.1	95.8	95.2	2.864	2.972	2.922

Table 3. Recognition Rates and Training Time of three databases - Method II

Wavelet Name	Recognition Rate (%)			Training Time in Seconds		
	ORL	JAFFE	ESSEX	ORL	JAFFE	ESSEX
Haar	98	97.8	97.6	1.559	1.685	1.799
Sym4	97.3	96.3	96.1	2.167	2.861	2.618
Sym8	96.67	96.5	96.4	2.182	3.132	2.732
Db4	97.04	97	95.7	2.227	2.664	2.932
Db6	96	96.4	97	2.212	2.792	2.658
Coif2	96.2	96	97.1	2.153	2.612	2.622
Coif4	95.5	95.2	95.6	2.249	2.853	2.714

Table 4. Comparison of Recognition Rates

Method Name	ORL	Method Name	JAFFE	Method Name	Essex
Eigen faces	89.5%	MLA+NN	91.14%	Wavelet+HMM	84.2%
Direct LDA	90.8%	LDA+SVM	91.27%	DWT+PCA	86.1%
Eigenfaces+NN	91.2%	SVM	91.6%	PZM	88.02%
Waveletface+NN	93.5%	Adaboost	92.4%	Gabor+SHMM	88.7%
Gabor+rank corr.	96%	LBP +SVM	92.0%	DM	91.72%
2DPCA	96%	PCA+SVM	93.43%	Fisher faces	92.62%
Push Pull MDA	97.54%	MLA + NM	97%	Curvelet+SVM	97.2%
KNN	90.33%	KNN	89.67%	KNN	89.5%
Method - I	98.3%	Method - I	98.0%	Method - I	98.0%
Method - II	98%	Method - II	97.8%	Method - II	97.6%

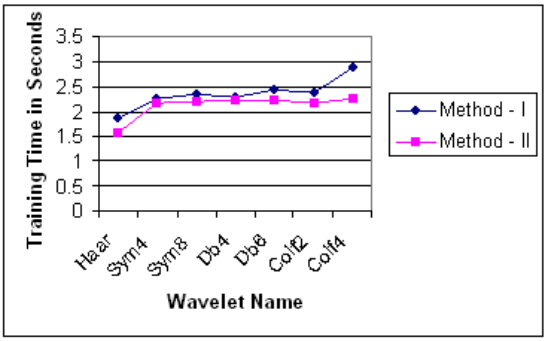


Fig. 6. Comparison of training time of ORL database

5 Conclusion

A face recognition system is devised with effective face representation by the collective approach of averaged wavelet or averaged wavelet packet transformation with discriminative common vectors and LBP method. The generated histogram values of the LBP are recognized by RBF network. The recognition rate obtained using method-I is greater than the proposed method-II. But the training time is less for method-II compared to method-I. The recognition rate is not affected by the collective method used in the preprocessing stage. The proposed system reduces the number of features hence minimizes the computational complexity and yielded the better recognition rates for ORL database, JAFFE face database and ESSEX database. The recognition performance of the classification is improved due to the combined technique used in the preprocessing stage which provides necessary discriminate information for classification and RBF network.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs Fisher faces: Recognition using class specific linear projection. IEEE Trans. Pattern Analysis Machine Intelligence 20(7), 711–720 (1997)
3. Carlos, M.T., Marcos, D.P., Miguel, A.F., Jesus, B.A.: Reducing Features using Discriminative Common Vectors. Cognitive Computation 2, 160–164 (2010)
4. Cevikalp, H., Neamtu, M., Wilkes, M.: Discriminative common vectors method with kernels. IEEE Trans. Neural Network 17(6), 1550–1565 (2006)
5. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Trans. Pattern Analysis Machine Intelligence 27(1), 4–13 (2005)
6. Er, M.J., Wu, S., Lu, J., Toh, H.L.: Face Recognition with Radial Basis Function (RBF) Neural Networks. IEEE Transactions on Neural Networks 13(3), 697–710 (2002)

7. Feng, G.C., Yuen, P.C., Dai, D.Q.: Human face recognition using PCA on wavelet subband. J. Electron. Imaging 9, 226–233 (2001)
8. Garcia, C., Zikos, G., Tziritas, G.: Wavelet packet analysis for face recognition. Image and Vision Computing 18, 289–297 (2000)
9. Jing, X.Y., Yao, Y.F., Yang, J.Y., Zhang, D.: A novel face recognition approach based on kernel discriminative common vectors (KDCV) feature extraction and RBF neural network. Neurocomputing 71, 3044–3048 (2008)
10. Kathirvalavakumar, T., Vasanthi, J.J.B.: Face representation using Wavelet, DCV and Modified Local Binary Patterns and Recognition by RBF. Journal of Machine Learning and Cybernetics (2013)
11. Li, B., Yin, H.: Face Recognition Using RBF Neural Networks and Wavelet Transform. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) ISNN 2005. LNCS, vol. 3497, pp. 105–111. Springer, Heidelberg (2005)
12. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Analysis Machine Intelligence 24(7), 971–987 (2002)
13. Pujol, A.F., Garca, J.C.: Computing the Principal Local Binary Patterns for face recognition using data mining tools. Expert Systems with Applications 39(8), 7165–7172 (2012)
14. Swets, D.L., Weng, J.: Using Discriminant Eigen features for Image Retrieval. IEEE Trans. Pattern Analysis and Machine Intelligence 18(8), 831–836 (1996)
15. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. 3(7), 71–86 (1991)
16. Perlibakas, V.: Face Recognition Using Principal Component Analysis and Wavelet Packet Decomposition. INFORMATICA 15(2), 243–250 (2004)
17. Wen, Y.: An improved discriminative common vectors and support vector machine based face recognition approach. Expert Systems with Applications 39(4), 4628–4632 (2012)
18. Wong, Y.W., Seng, K.P., Ang, L.M.: Dual optimal multiband features for face recognition. Expert Systems with Applications 37(4), 2957–2962 (2010)
19. Wong, Y.W.: Radial Basis Function Neural Network with Incremental Learning for Face Recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(4), 940–949 (2011)
20. Zhang, B.L., Zhang, H., Ge, S.S.: Face Recognition by Applying Wavelet Subband Representation and Kernel Associative Memory. IEEE Transactions on Neural networks 15(1), 166–177 (2005)
21. Zhou, S.R., Yin, J.P., Zhang, J.M.: Local binary pattern (LBP) and local phase quantization (LBQ) based on Gabor filter for face representation. Neurocomputing 116(20), 260–264 (2013)

Face Recognition in Very Low Bit Rate SPIHT Compressed Facial Images

Karuppana Gounder Somasundaram[1] and Nagappan Palaniappan[2]

[1] Image Processing Lab, Department of Computer Science and Applications, The Gandhigram Rural Institute Deemed University, Gandhigram
ka.somasundaram@gmail.com
[2] Computer Centre,The Gandhigram Rural Institute Deemed University, Gandhigram
naapala@gmail.com

Abstract. Biometric authentication procedures use facial images for authentication in an organization. Storing of large facial images require large storage space. To reduce the storage space, compression methods are employed. JPEG and JPEG2000 compression standards are recommended by ISO standard for facial recognition format up to the file size of 12 kB. In this paper we propose a scheme to evaluate the facial recognition performance of the very low bit rate facial images. The facial images are compressed using SPIHT encoding. The proposed scheme makes use of PCA and ICA methods to evaluate the results obtained. The very low bit rate facial images perform equivalent to the higher bit rate images in facial recognition.. ...

Keywords: computational geometry, graph theory, Hamilton cycles.

1 Introduction

In digital image processing biometric is the process of storing and analyzing the characteristics of human organs and behavior like finger print, iris patterns, retinal images, palm print, signature, hand writing, voice and facial images for personal identification. Compared with other biometric characteristics of human organs, the face plays the key role. The face is not only one of the first visual patterns infants learn to recognize, but it is also one of the natural means for human beings to recognize each other [1]. Automated Facial Recognition (AFR) is used in verifying or identifying persons. In recent days, smart cards with a low memory chip for storing the biometric data are used for automatic facial recognition. Thousands of facial images are to be stored in data servers for automating identification or verification process. During this process, the clients will query the server to transmit a particular image over a network. When the size of the facial image is large, undue delay will be caused in the identification. Even it requires data servers with high storage capacity to store and a lot of bandwidth to transmit. At this stage, automatic facial recognition methods based on compressed images are required.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 160–172, 2013.
© Springer International Publishing Switzerland 2013

AFR methods require normalization and compression of facial images. The ISO Standard for Face Recognition Format for Data Interchange [2] states that, "The full frontal face image type, including the full head, neck and shoulders is being proposed for Machine Readable Travelling Documents (MRTD). The image data shall be encoded using either JPEG or JPEG2000 and that compression to 0.44 bits per pixel, corresponding to an approximate file size of 12 kB, may be a good target".

Most of the previous works are based on JPEG compression algorithm in pixel domain and a very few of them are dealing with JPEG2000 in pixel and transform domain. Blackburn et al., [3] in their Facial Recognition Vendor Test (FRVT) 2000 Evaluation report verified the effect of JPEG image compression technique with bits per pixel (bpp) 0.8, 0.4, 0.25 and 0.2 on FERET database. They concluded that JPEG compression of facial images does not affect recognition performance and even the performance slightly increased for some compression rates. Compression below 0.2 bpp the recognition performance is dropped significantly. Wijaya et al., [4] concluded that the correlation filters achieve higher recognition rates with JPEG2000 compressed images at the rate of 0.5 bpp. McGarry et al., [5] verified the effects of JPEG2000 compression and found out that the compression rates higher than 10:1 (0.8 bpp) will not produce any performance drop. Delac et al., [6,7,8] studied the effects of image compression in facial recognition using JPEG and JPEG2000 on FERET dataset with PCA, ICA and LDA algorithms at the bit rates of 1.0, 0.5, 0.3 and 0.2. They also concluded that the effects of JPEG and JPEG2000 compression does not deteriorate recognition rate significantly and even the recognition rate is slightly improved when using compressed facial images. Mustafa Ersel Kamasak and Bulent Sankur [9] made an attempt to test the three compression schemes for facial recognition - Vector Quantization, JPEG and Set Partition in Hierarchical Coding (SPIHT) [10] methods. They found that SPIHT is the most robust compression scheme among them. When comparing all the above works, it is very clear that the JPEG or JPEG2000 compression is only done at the $bpp \leq 0.2$.

The aim of this paper is to evaluate the AFR performance of the very low bit rate compressed images (below 0.2 bpp). The facial images are compressed using SPIHT encoding at the bit rates 1.0 bpp, 0.5 bpp, 0.3 bpp, 0.2 bpp, 0.1 bpp, 0.08 bpp and 0.05 bpp. The obtained results are compared with the AFR performance of the JPEG2000 compressed images. To evaluate the performance Principal Component Algorithm (PCA) [11] and Independent Component Algorithm (ICA) [12] are involved and facial images from the standard data set for facial recognition FERET [13] are used. All the images used for testing the recognition performance are normalized as per the ISO/IEC 19794-5 Standards [2] for facial image format before compression.

The remaining of the paper is organized as follows. Section 2 explains the various methods involved. The proposed method is detailed in section 3. Results and analysis is presented in the section 4 and section 5 concludes the paper.

2 Introduction

2.1 Facial Image Normalization

As per ISO/IEC 19794-5 Standards [2], the fundamental facial image types are full frontal and token frontal images. Full frontal image is the facial image which includes full head with all hair, neck and shoulders. Token frontal image is the full frontal image with specific dimensions. Eye positions are fixed. The inter-pupillary distance (distance between the eyes) is based on the dimension of the image. The facial images are illumination normalized using histogram equalization. The eye centers are located using face detection algorithm [14]. Using the eye centers, head roll angle is calculated and the image is rotated to eliminate it. Inter pupillary distance is estimated and adjusted to 60 pixels. The core facial area is adjusted to 128 x 128 pixels for our proposed scheme.

2.2 Feret Database

FERET database is the standard database for testing facial recognition algorithms. This database is collected by Defense Advance Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST) of United States of America (USA) from 1993 to 1997. The total collection counts to 14051 grayscale facial images. Images are categorized into various groups depending upon the nature as Fa, Fb, Fc, Dup I and Dup II with 1196, 1195, 194, 722 and 234 images respectively.

Gallery images are the collection of facial images from known individuals which forms the search dataset. Probe images are the collection facial images of unknown persons to be recognized by matching.

2.3 SPIHT Coding

SPIHT is a simple and efficient algorithm for encoding wavelet coefficients. The transmission of embedded bit stream can be stopped at any bit rate and the required image can be reconstructed. This method encodes the image as a single tile which avoids blocking effects even at very low bit rates giving good rate-distortion performance. SPIHT encoding is based on spatial oriented trees (SOT).

Three lists, list of insignificant pixels (LIP), list of insignificant sets (LIS) and list of significant pixels (LSP) are used. There are two passes, sorting pass and refinement pass. This method requires N+1 iterations, where N=$log_2(\max(W_{(i,j)}))$. i,j represent the i^{th} row and j^{th} column of co-efficient matrix W obtained by wavelet transform. The iterations are processed with the thresholds 2^n, 2^{n-1}, 2^{n-2},..2^1, 2^0.

In the sorting pass, the wavelet coefficients in LIP are tested for significance. In every iteration the coefficients which satisfy the condition, $2^n <$abs$(W_{(i,j)}) \leq 2^{n+1}$ become significant else insignificant. The significant coefficients from LIP are added to LSP. Sets in LIS are sequentially tested. Significant sets are

partitioned and added to LIS, LIP and LSP as new subsets. When a coefficient is significant a bit value 1, otherwise 0 is sent to the decoder along with a sign bit 0 for positive and 1 for negative coefficients. Instead of comparing all the coefficients to reduce the complexity, the spatial oriented trees are checked by performing the test: abs(Max($W_{(i,j)}$) $\in T_k$) $>2^n$ where, Tk denotes the kth SOT and $W_{(i,j)}$ belongs to the spatial oriented tree T_k. If this condition fails, relative SOT is skipped in that particular iteration. The refinement pass sends one bit of the coefficients in LSP generated in the previous iterations from the MSB to the LSB.

2.4 PCA and ICA

PCA. An applied linear algebra tool used for dimensionality reduction of the given data set. It decorrelates the second-order statistics of the data. A 2-D facial image is converted into a single dimensional vector by joining all the rows one after another having r (row) x c (columns) elements. For M training images, there will be M single dimensional vectors. A mean centered image is calculated by subtracting the mean image from each vector. Based on the covariance matrix of the mean centered image, eigen vectors are computed. The basis vectors which represent the maximum variance direction from the original image are selected as feature vectors. These feature vectors are named as eigen faces or face space. It is not necessary that the number of feature vectors should be equal to the number of training images. Every image in the gallery image set is projected into the face space and the weights are stored in the memory. The face to be probed is also projected into the face space. The distance between the projected probe image weights and every projected gallery image weight is computed. The gallery image having the shortest distance will be treated as the recognized face.

ICA. The next generation of PCA, which makes the signals independent to the maximum by decorrelating the second-order statistics and reducing higher-order statistic dependencies. The resultant independent basis vectors form the face space. Bartlett et al., [12] provided two different architectures Architecture - I which treated images as random variables and the pixels as outcomes and Architecture II treated the pixels as random variables and the images as outcomes. This ICA was derived from the principal of optimal information transfer through sigmoidal neurons. The dimensionality of the basis vectors are reduced using PCA. ICA Architecture I is used in our scheme.

2.5 Distance Measures

The distance measures are used to compare the similarity between two feature vectors. The distance measures used in our work are L1, L2 and cosine angle. Let x and y are two vectors of size n and d is the distance between the vectors x and y.

L1 Distance. City-Block or Manhattan distance is defined as the sum of the absolute differences between the two vectors

$$d(x, y) =\mid x - y \mid= \sum_{i=1}^{n} \mid x_i - y_i \mid \tag{1}$$

L2 Distance. Euclidean distance is defined as the sum of the squares of the differences

$$d(x, y) = \|x - y\|^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{2}$$

Cosine Angle. It is the angle between the feature vectors. The scalar product of the two vectors is divided by their magnitudes.

$$d(x, y) = -\frac{x.y}{\|x\|\|y\|} = -\frac{\sum_{i=1}^{n} x_i y_i}{[\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i]^{\frac{1}{2}}} \tag{3}$$

2.6 Performance Measures

Based on the previous works [8,15] the performance measures Recognition Rate (RR) and Normalized Recognized Rate (NRR) are defined.

Recognition Rate (RR). Let P be the probe set with n images and G be the gallery set with k images. It is assumed that for every probe image P_n there is a gallery image G_k. Both probe P and gallery G images are projected in the face space. For every probe image P_n, the similarity scores S(G_k) of all the gallery images in G are calculated using the distance measures and sorted with decreasing similarity.

$$S(G_k) = distance(P_n, G_k), k = 1 \tag{4}$$

The smallest score S(G_k) of the gallery image G_k is considered to be the closest match to the probe image P_n. If the gallery image Gk is exactly equal to the probe image P_n, then it can be declared that the algorithm correctly recognized the probe image. For n probe images, if R_n images are correctly recognized than the recognition rate is defined as the ratio between the number of correctly recognized and the total number of probe images R_n/n. For example, out of 1195 probe images 1011 are correctly recognized than the recognition rate is 1011/1195 = 84.60.

Normalized Recognition Rate (NRR). Normalized recognition rate is used to compare the recognition rates in uncompressed and compressed domain. Both are calculated with the same environment. Let RR_C be the recognition rate in

the compressed domain and RR_U be the recognition rate in the uncompressed domain, then NRR = RR_C/RR_U. If NRR >1 then the recognition rate is higher in compressed domain than the uncompressed domain. If NRR <1 then the recognition rate is less in compressed domain than the uncompressed domain. NRR unambiguously shows in which domain the higher recognition rate can be achieved for the particular bit rate of compression.

3 Proposed Method

The normalized images from FERET database are compressed to the low bit rates 1.0, 0.5, 0.3 and 0.2 bpp as used in the previous works. To compress the facial images with JPEG2K standard Kakadu [16] or Jasper [17] software are used. For SPIHT compression Matlab version of SPIHT coding is used. Fig.1 shows the images compressed at the bit rates 1.0 bpp, 0.5 bpp, 0.3 bpp and 0.2 bpp using JPEG2K and SPIHT compression algorithms.

Fig. 1. Low bit rate compressed images. JPEG2000 (i) 1.0 bpp-2kB (ii) 0.5 bpp-1 kB (iii) 0.3 bpp-0.6 kB (iv) 0.2 bpp-0.4 kB, SPIHT (v) 1.0 bpp-2kB (vi) 0.5 bpp-1 kB (vii) 0.3 bpp-0.6 kB (viii) 0.2 bpp-0.4 kB

To test the performance of facial recognition in very low bit rates, compression rates 1.0 bpp, 0.08 bpp and 0.05 bpp are applied. When facial images are compressed at these bit rates below 0.2 bpp (less than 204 bytes), JPEG2K-Kakadu software generates just a gray scale image of uniform pixel values. When tried with JPEG2K-Jaspar software it is possible to compress the images to the required bit rates, but the resulting file sizes are very high. An image of size 128 x 128 pixels compressed at the rate of 0.05 bpp results in a file size of 802 bytes instead of 103 bytes. So, only the SPIHT compressed images are used to verify the AFR performance. Fig.2 shows the images compressed at very low bit rates 0.1 bpp, 0.08 bpp and 0.05 bpp.

In all our experiments the training and gallery images are uncompressed and only the probe images are compressed. The FERET image set Fa is used as gallery image set. Sets Fb, Fc, Dup I and Dup II are used as probe image sets. A training set of 501 images from FERET data set obtained from the CSU Face Identification Evaluation System of Colorado State University [18] is used in our experiment. Among these training images 80% are from gallery images and 20% from Dup I images. While performing PCA on the training set, it generates 500 eigen vectors. Among these 500 eigen vectors only the top 200 eigen vectors (40%

Fig. 2. Very low bit rate compressed images (i) 0.1 bpp-0.2 kB (ii) 0.08 bpp-0.16 kB (iii) 0.05 bpp-0.1 kB

of the total eigen vectors) are selected as basis vectors. These basis vectors are used with PCA and ICA algorithms to generate the PCA face space (W_{PCA}) and the ICA face space (W_{ICA}).

Initially the W_{PCA} and W_{ICA} face spaces are generated and stored using the uncompressed training images. The uncompressed gallery images are projected into the W_{PCA} and W_{ICA} face spaces and its resultant projected weights G_{PCA} and G_{PCA} are stored. Then the experiments were carried out with Fb, Fc, Dup I and Dup II probe sets.

For every probe set group, the JPEG2K and SPIHT compressed images are projected into the W_{PCA} and W_{ICA} face spaces resulting four projections P_{PJ}, P_{PS}, P_{IJ} and P_{IS}. The distance measures L1, L2 and COS are used to find the closest match from the gallery projections G_{PCA} and G_{ICA}. Based on the closest match, the recognition rates RR_{L1-PJ}, RR_{L1-PS}, RR_{L1-IJ}, RR_{L1-IS}, RR_{L2-PJ}, RR_{L2-PS}, RR_{L2-IJ}, RR_{L2-IS}, RR_{COS-PJ}, RR_{COS-PS}, RR_{COS-IJ}, and RR_{COS-IS} are calculated. The recognition rate (RR_{UC}) is calculated using the uncompressed probe images. Using RR_{UC} the normalized recognition rates NRR_{L1-P}, NRR_{L1-I}, NRR_{L2-P}, NRR_{L2-I}, NRR_{COS-P}, NRR_{COS-I} are also calculated.

4 Results and Analysis

The calculated RR and NRR for every probe group Fb, Fc, Dup I and Dup II are listed in the following tables individually. The third row in every table heading shows the AFR rate of uncompressed images.

Table 1. Recognition Rates of Fb images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	83.93		89.96		83.77		87.45		83.51		89.04	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	82.76	82.76	90.06	90.13	83.26	82.93	87.53	87.45	82.85	82.85	89.12	88.95
0.5	83.18	82.93	90.46	90.76	83.35	83.35	87.95	87.53	82.85	82.85	89.04	88.95
0.3	82.85	82.85	90.54	89.96	83.43	83.35	88.37	88.03	83.10	83.18	89.04	89.04
0.2	82.93	82.99	91.21	91.05	83.01	83.35	88.87	88.20	83.43	82.76	89.12	89.04
0.1	0	83.01	0	90.88	0	82.85	0	88.87	0	83.26	0	89.04
0.08	0	83.10	0	90.96	0	83.51	0	88.95	0	83.26	0	88.79
0.05	0	81.84	0	91.05	0	81.76	0	88.61	0	81.42	0	88.70

Table 1 shows the recognition rates (RR) of the Fb images. For the probe set Fb, the RR of ICA method shows significant increase than the PCA method for all the cases. L2 gives better RR when compared with L1 and COS for PCA methods. L1 gives higher results than L2 and COS in ICA methods. Both JPEG2K and SPIHT based compression methods achieve RR with a marginal difference of ±1%. No significant higher difference is seen in the results. For the bit rate of 1.0 bpp, RR_{L1-IS} gets the higher RR of 90.13, for 0.5 bpp again RR_{L1-IS} gets 90.76 as higher RR. RR_{L1-IJ} achieves a top RR of 90.54 for 0.3 bpp. For 0.2 bpp, again RR_{L1-IJ} gets the top RR of 91.21.

RR_{L1-IS} gets the higher RR for the bit rates 0.1 bpp, 0.08 bpp and 0.05 bpp with RR 90.88, 90.96 and 91.05 respectively. While considering the bit rates 0.1 bpp and 0.08 bpp RR is more or less equal to the RR_{UC} for PCA methods. The bit rate 0.05 bpp drops significantly when compared with all other RR for the PCA methods. Considering ICA methods all the bit rates 0.1 bpp, 0.08 bpp and 0.05 bpp gives RR equivalent to other bit rates. Especially RR_{L1-IS} gives the top RR 91.05 at the bit rate of 0.05 and RR_{L2-IS} gives the highest RR 88.95 at the bit rate of 0.08. On overall comparison RR_{L1-IJ} achieves the highest RR of 91.21. When compared with all the RR the compressed images give higher RR than the uncompressed images.

Table 2. Normalized Recognition Rates of Fb images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	0.986	0.986	1.001	1.002	0.994	0.990	1.001	1.000	0.992	0.992	1.001	0.999
0.5	0.991	0.988	1.006	1.009	0.995	0.995	1.006	1.001	0.992	0.992	1.000	0.999
0.3	0.987	0.987	1.006	1.000	0.996	0.995	1.011	1.007	0.995	0.996	1.000	1.000
0.2	0.988	0.989	1.014	1.012	0.991	0.995	1.016	1.009	0.999	0.991	1.001	1.000
0.1	- 0.989		- 1.010		- 0.989		- 1.016		- 0.997		- 1.000	
0.08	- 0.989		- 1.011		- 0.997		- 1.017		- 0.997		- 0.997	
0.05	- 0.975		- 1.012		- 0.976		- 1.013		- 0.975		- 0.996	

Normalized recognition rates of Fb images are given in table 2. Considering the NRR, all the ICA methods attain the value more than 1 except RR_{COS-IS}. RR_{L2-IS} achieves the highest NRR of 1.017 at the bit rate of 0.08 bpp. All the PCA methods have NRR less than 1. It clearly shows that the ICA methods give better RR then the PCA methods in compressed domain.

Table 3 shows the recognition rates of Fc images. Experimenting with the probe set Fc, the RR of ICA method shows significant increase than the PCA method for L2 and COS and for L1 it is lower than PCA for higher bit rates. L1 gives better RR when compared with L2 and COS for PCA methods. The COS gives higher results than L1 and L2 for ICA methods. Both JPEG2K and SPIHT achieve RR with a marginal difference of ±2 %. The RR for PCA based L2 and COS are very low, even below 50% when compared with $L1_{PCA}$ and all ICA based distance measures. For the bit rate of 1.0 bpp, RR_{COS-IJ} and

Table 3. Recognition Rates of Fc images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	32.99		28.87		16.49		29.90		17.01		35.57	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	33.51	32.99	28.35	28.87	16.49	16.49	29.38	30.41	16.49	16.50	34.54	34.54
0.5	34.02	31.44	31.44	28.87	15.98	16.49	29.90	29.90	16.49	16.49	35.57	36.60
0.3	32.99	33.51	32.47	31.96	16.49	16.49	34.02	31.44	15.98	17.01	36.60	35.57
0.2	32.47	32.47	34.54	32.47	15.46	16.49	36.08	31.96	16.49	15.98	37.63	37.11
0.1	0	31.44	0	34.54	0	15.46	0	37.63	0	16.49	0	38.14
0.08	0	29.90	0	36.60	0	14.95	0	37.73	0	16.49	0	37.11
0.05	0	27.84	0	37.63	0	15.46	0	38.14	0	15.99	0	38.14

RR_{COS-IS} gets the higher RR of 34.54, for 0.5 bpp again RR_{COS-IS} gets 36.60 as higher RR. RR_{COS-IJ} achieves a higher RR of 36.60 and 37.63 for 0.3 bpp and 0.2 respectively.

Table 4. Normalized Recognition Rates of Fc images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	1.016	1.000	0.982	1.000	1.000	1.000	0.983	1.017	0.969	0.970	0.971	0.971
0.5	1.031	0.953	1.089	1.000	0.969	1.000	1.000	1.000	0.969	0.969	1.000	1.029
0.3	1.000	1.016	1.125	1.107	1.000	1.000	1.138	1.052	0.939	1.000	1.029	1.000
0.2	0.984	0.984	1.196	1.125	0.938	1.000	1.207	1.069	0.969	0.939	1.058	1.043
0.1	-	0.953	-	1.196	-	0.938	-	1.259	-	0.969	-	1.072
0.08	-	0.906	-	1.268	-	0.907	-	1.262	-	0.969	-	1.043
0.05	-	0.844	-	1.303	-	0.938	-	1.276	-	0.969	-	1.072

RR_{COS-IS} gets the higher RR of 38.14 for the bit rate 0.1 bpp. The bit rate 0.08 bpp attains its top RR 37.73 with RR_{L2-IS}. Both RR_{L2-IS} and RR_{COS-IS} achieves their top RR of 38.14 for 0.05 bpp. While considering the bit rates 0.1 bpp, 0.08 bpp and 0.05, the RR is more than the RR_{UC} for ICA methods. In this case of Fc images, the RR of bit rate 0.05 bpp is higher than all other bit rates in the case of ICA methods and its drops significantly for PCA methods. When compared with all the methods and bit rates RR_{L2-IS} at 0.05 bpp, RR_{COS-IS} at 0.05 bpp and RR_{COS-IS} at 0.1 bpp achieves the highest RR of 38.14. Based on these results, it can be assumed that the effects of illumination variations in the Fc group images are eliminated in the case of low bit rate compressed images.

Table 4 shows the normalized RR of Fc images. For all of the distance measures, ICA-SP methods achieve NRR higher than 1 except for NRR_{COS-IS} at 1.0 bpp. The highest NRR value obtained 1.072 is achieved for NRR_{COS-IS} for 0.1 bpp and 0.05 bpp. ICA-J2K methods achieve NRR less than 1 for all the distance measures at the bit rate 1.0 bpp. When considering PCA methods,

most of the NRR values are less than 1. Especially in the case of low bit rates the PCA-SP NRR values are below 1. From these results we can assume that the RR of PCA methods is better in uncompressed images and ICA methods is better in compressed images.

Table 5. Recognition Rates of Dup I images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	36.98		39.75		35.19		36.84		36.43		37.95	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	37.81	38.37	39.61	39.20	35.46	35.45	36.70	36.55	36.01	35.87	38.64	38.09
0.5	38.37	38.37	39.06	39.33	35.46	35.32	36.84	36.43	36.14	36.43	38.64	37.67
0.3	38.50	38.37	39.20	39.06	35.73	35.73	36.57	36.73	36.01	36.29	39.20	38.92
0.2	36.43	37.81	39.61	40.03	34.76	35.46	37.40	37.26	35.46	36.43	37.81	38.05
0.1	0	37.53	0	39.75	0	34.76	0	37.40	0	35.46	0	38.50
0.08	0	35.73	0	40.72	0	34.76	0	37.67	0	35.87	0	39.06
0.05	0	33.24	0	39.89	0	33.38	0	37.67	0	33.93	0	38.78

RR of Dup I group images are given in table 5. For this image, the RR of ICA method shows significant improvement than the PCA method. L1 gives better RR when compared with L2 and COS for PCA and ICA methods. Both JPEG2K and SPIHT based compression methods achieve RR with a marginal difference of ±1%. For the bit rate of 1.0 bpp, RR_{L1-IJ} gets the higher RR of 39.20, for 0.5 bpp again RR_{L1-IS} gets 39.33 as higher RR. RR_{L1-IJ} and RR_{COS-IJ} achieves a top RR of 39.20 for 0.3 bpp. For 0.2 bpp, again RR_{L1-IS} gets the higher RR of 40.03.

RR_{L1-IS} gets the higher RR of 39.75 for the bit rate 0.1 bpp. The bit rate 0.08 bpp attains its higher RR 40.72 with RR_{L1-IS}. RR_{L1-IS} achieves its higher RR of 39.89 for 0.05 bpp. While considering the bit rates 0.1 bpp, 0.08 bpp and 0.05, all the RR is more than the RR_{UC} for ICA methods. For Dup I images, the RR of bit rate 0.08 bpp is the highest than all other bit rates. But for PCA methods it drops significantly with the bit rate. When compared with other methods and bit rates, RR_{L1-IS} at 0.08 bpp achieves the highest RR of 40.72. The Dup I group low bit rate compressed images can be well recognized by ICA method L1, L2 and COS.

Table 6 shows the NRR of Dup I images. The NRR of ICA methods are more than 1 for very low bit rates and less than 1 for low bit rates for all the distance measures. But the PCA methods achieve higher NRR for low bit rates and very lower NRR for very low bit rates. The NRR of ICA and PCA methods are inversely proportional to each other.

Table 7 shows the recognition rates of Dup II images. While using the probe set Dup II also the RR of ICA method shows significant improvement than the PCA method. L1 achieves higher RR when compared with L2 and COS for PCA and ICA methods. Both JPEG2000 and SPIHT methods achieve RR with

Table 6. Normalized Recognition Rates of Dup I images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	1.022	1.038	0.996	0.986	1.008	1.008	0.996	0.992	0.988	0.985	1.018	1.004
0.5	1.038	1.038	0.983	0.989	1.008	1.004	1.000	0.989	0.992	1.000	1.018	0.993
0.3	1.041	1.038	0.986	0.983	1.015	1.015	0.993	0.997	0.988	0.996	1.033	1.026
0.2	0.985	1.022	0.996	1.007	0.988	0.988	1.015	1.011	0.973	1.000	0.996	1.003
0.1	- 1.015		- 1.000		- 0.988		- 1.015		- 0.973		- 1.014	
0.08	- 0.966		- 1.024		- 0.988		- 1.023		- 0.985		- 1.029	
0.05	- 0.899		- 1.004		- 0.949		- 1.023		- 0.931		- 1.022	

Table 7. Recognition Rates of Dup II images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	13.24		20.09		10.68		17.52		11.97		15.38	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	14.96	16.24	20.09	20.09	11.97	11.97	17.52	17.09	12.82	12.39	16.24	15.81
0.5	16.24	15.34	20.09	19.23	11.54	11.97	17.52	17.09	12.39	12.39	16.24	14.96
0.3	15.81	15.38	20.09	19.66	11.54	11.97	17.52	17.52	12.39	11.97	16.24	15.81
0.2	13.68	14.53	20.09	21.37	10.21	11.11	18.38	17.95	11.11	11.96	15.38	16.24
0.1	0	14.53	0	21.37	0	10.68	0	17.09	0	11.54	0	15.81
0.08	0	13.25	0	22.65	0	10.21	0	17.95	0	11.96	0	16.24
0.05	0	11.54	0	21.75	0	09.83	0	18.38	0	9.83	0	15.81

a marginal difference of ±1.5%. The RR of PCA method drops along with bit rates except the RR_{L1-PJ} at 0.5 bpp. For the bit rate of 1.0 bpp, RR_{L1-IJ} and RR_{L1-IS} gets the higher RR of 20.09, for 0.5 bpp RR_{L1-IJ} gets 20.09 as higher RR. RR_{L1-IJ} achieves a top RR of 20.09 for 0.3 bpp. For 0.2 bpp, again RR_{L1-IS} gets the top RR of 21.37. RR_{L1-IS} gets the higher RR of 21.37 for the bit rate 0.1 bpp. The bit rate 0.08 bpp attains its higher RR 22.65 with RR_{L1-IS}. RR_{L1-IS} achieves its higher RR of 21.75 for 0.05 bpp. While considering the bit rates 0.1 bpp, 0.08 bpp and 0.05, all the RRs are more than the RR_{UC} for ICA methods. For Dup II images, the RR of bit rate 0.08 bpp is the highest than all other bit rates. When compared with all the methods and bit rates RR_{L1-IS} at 0.08 bpp achieves the highest RR of 22.65. The low bit rate SPIHT compressed Dup II images perform better in facial recognition.

Normalized recognition rates of Dup II images are shown in table 8. All the ICA methods attain the NRR value higher than 1 except very few cases. Both PCA-JPEG2K methods and PCA-SPIHT methods achieve more or less equal NRR. The highest NRR is achieved by RR_{L1-PS}.

Considering the overall performance ICA performs better than PCA for both JPEG2K and SPIHT compressed images. L1 outperforms L2 and COS for the types Fb, Dup I and Dup II. The COS performs well for Fc images. While verifying the RR, JPEG2K compressed images performs better for Fb and Fc

Table 8. Normalized Recognition Rates of Dup II images

BPP	L1				L2				COS			
	PCA		ICA		PCA		ICA		PCA		ICA	
	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP	J2K	SP
1.0	1.130	1.227	1.000	1.000	1.121	1.121	1.000	0.975	1.071	1.035	1.056	1.028
0.5	1.227	1.159	1.000	0.957	1.081	1.121	1.000	0.975	1.035	1.035	1.056	0.973
0.3	1.194	1.162	1.000	0.979	1.081	1.121	1.000	1.000	1.035	1.000	1.056	1.028
0.2	1.033	1.097	1.000	1.064	0.956	1.040	1.049	1.025	0.928	0.999	1.000	1.056
0.1	0	1.097	0	1.064	0	1.000	0	0.975	0	0.964	0	1.028
0.08	0	1.001	0	1.27	0	0.956	0	1.025	0	0.999	0	1.056
0.05	0	0.872	0	1.083	0	0.920	0	1.049	0	0.821	0	1.028

images and SPIHT compressed images performs well for Dup I and Dup II images. The low bit rate SPIHT compressed images attains higher RR than the higher bit rate compressed images using ICA method.

5 Conclusion

Our scheme evaluates the AFR performance of the very low bit rate images compressed using SPIHT encoding method using PCA and ICA methods. The SPIHT compressed images attain RR equal to the JPEG2000 compressed images. Especially in the case of very low bit rates (0.1, 0.08 and 0.05 bpp), ICA methods perform better than the low bit rate (1.0, 0.5, 0.3 and 0.2 bpp) compressed images and tolerant to various illuminations.

References

1. Jain, A.K.: Biometric Recognition: How Do I know Who You Are? In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 19–26. Springer, Heidelberg (2005)
2. Biometric Data Interchange Formats Part 5: Face Image Data draft revision 19794-5 (2004)
3. Blackburn, D.M., Bone, J.M., Philips, P.J.: FRVT 2000 Evaluation Report (2001), `http://www.frvt.org/FRVT2000/documents.htm`
4. Wijaya, S.L., Savvides, M., Vijaya Kumar, B.V.K.: Illumination-Tolerant Face Verification of Low-Bit-Rate JPEG2000 Wavelet Images with Advanced Correlation Filters for Handheld Devices. Journal fo Applied Optics 44, 655–665 (2005)
5. McGarry, D.P., Arndt, C.M., McCabe, S.A., D'Amato, D.P.: Effects of compression and individual variability on face recognition performance. In: Proc. of SPIE, vol. 5404, pp. 362–372 (2004)
6. Delac, K., Grgic, M., Grgic, S.: Effects of JPEG and JPEG2000 Compression on Face Recognition. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, vol. 3687, pp. 136–145. Springer, Heidelberg (2005)
7. Delac, K., Grgic, M., Grgic, S.: Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set. International Journal of Imaging Systems and Technology, 252–260 (2006)

8. Delac, K., Grgic, M., Grgic, S.: Face Recognition in JPEG and JPEG2000 Compressed Domain. Image and Vision Computing 27, 1108–1120 (2009)
9. Kamasak, M.E., Sankur, B.: Face Recognition under lossy Compression. In: The Fifth International Conference on Pattern Recognition and Information Processing (1999)
10. Said, A., Pearlman, W.A.: A New, Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees. IEEE Transaction Circuit Systems and Video Technology 6, 243–249 (1996)
11. Turk, M.A., Pentland, A.P.: Face Recognition using Eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–591 (1991)
12. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face Recognition by Independent Component Analysis. IEEE Transactions on Neural Networks 14, 1450–1464 (2002)
13. Grayscale FERET Database, http://www.itl.nist.gov/iad/humanid/feret/
14. Somasundaram, K., Palaniappan, N.: Personal ID Image Normalization Using ISO/IEC 19794-5 Standards for Facial Recognition Improvement. In: Balasubramaniam, P., Uthayakumar, R. (eds.) ICMMSC 2012. CCIS, vol. 283, pp. 429–438. Springer, Heidelberg (2012)
15. Moon, H., Philips, P.J.: Computational and Performance aspects of PCA-based Face Recognition Algorithms. Perception 30, 303–321 (2001)
16. Kakadu Software. JPEG2000 Standard, Part 1- ISO/IEC 15444-1, http://www.kakadusoftware.com
17. Jasper Software. JPEG2000 Standard, Part 1- ISO/IEC 15444-1, http://www.ece.uvic.ca/~frodo/jasper/
18. Evaluation of Face Recognition Algorithms, Colorado State University, http://www.cs.colostate.edu/evalfacerec/index10.php

Modified Kittler and Illingworth's Thresholding for MRI Brain Image Segmentation

T. Kalaiselvi and P. Nagaraja

Image Processing Lab, Department of Computer Science and Applications,
Gandhigram Rural Institute–Deemed University, Gandhigram–624 302,
Dindigul, Tamil Nadu, India
{kalaivpd,pnagaraja02}@gmail.com

Abstract. This work is aimed to produce a robust thresholding method for segmenting the MRI brain images. A popular thresholding method commonly used in digital image segmentation is the Kittler and Illingworth's (MET) method because it improves the segmentation process effectively. It is easy to implement and works well with the general images. However, it fails to segment the MRI brain images. This paper proposed a method to modify the objective function of traditional MET method by including the total variance of given image and a weight parameter. This method gives the satisfactory results for the MRI brain images, while compared with other threshold methods and traditional MET method. The segmented images are compared by using the region non–uniformity (NU) parameter. The NU value of proposed work is very low while compared with the original and other existing methods. The MRI brain images are segmented by the proposed work have sub structural clarity for further processing.

Keywords: Image Segmentation, Kittler and Illingworth, MRI brain Images, Thresholding.

1 Introduction

Image processing is a most significant study of digital image analysis, which is used to obtain necessary information from them. In medical image analysis, segmentation technique is an indispensible step for image processing. Image segmentation is the process of partioning a digital image into multiple segments and separating out mutually exclusive homogeneous regions of interest. Segmentation is based on measurements taken from the image and might be gray level, color, texture, depth or motion [1] [2]. Nowadays, segmentation of brain image is the most significant task for clinical applications. MRI brain images are oriented to dissimilar from two goals, namely classifying tissues and anatomical structures. MRI brain image tissue classes are contains four regions, namely gray matter (GM), white matter (WM), cerebral spinal fluid (CSF) and background [3].

Segmentation algorithms are classified into two types: discontinuity and similarity. Discontinuity is the abrupt changes on the gray level of image and similarity is the partioning the image into similar pixels within the regions. A popular

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 173–179, 2013.
© Springer International Publishing Switzerland 2013

method for image segmentation is thresholding and it is based on the second category. The approach of thresholding is converting any higher scale image into binary image, it separates the pixels that are background and foreground. The specified value is called as threshold value. Thresholding techniques are classified into two categories: global and local threshold. The first category is a single threshold value used in the whole image. The second category is threshold value assigned to each pixel it belongs to background or foreground of pixel using information around the pixel. The importance of thresholding is take minimum storage space, fast and simpler processing. Several thresholding techniques are developed and used with various applications and medical image analysis[1,2,4].

Thresholding techniques are most-widely used in image analysis processing and machine vision industry. In recent years, many thresholding methods have been developed. The most often thresholding methods used are Otsu, Ridler and Calvard, Kittler and Illingworth's minimum error thresholding method (MET) [5] [6]. A comparative study and an overview of thresholding methods can be found in Sauvola and Pietikainen [7]. They proposed a new technique to document image binarization and used two algorithms in order to calculate a different threshold for each pixel. Otsu is very popular method for global thresholding. This method minimizing the weighted sum of within-class variance of foreground and background pixels to establish an optimum threshold. It is good for bimodal or multimodal distribution and it has some difficulties for processing images with unimodal distributions with some limitations [8]. Soharab Hossain Shaikh et al proposed an iterative partitioning method that produces good results for graphic documents [9]. Nikolaos and Dimitris proposed a thresholding algorithm for historical manuscripts with good results [10].

MET method provides good results of general gray images. Unfortunately, this thresholding not provide well segmented image in MRI brain images. MET method may not work properly in MRI brain images, because it is histogram and probability based method [11]. A survey of thresholding technique gave best rank for MET method [12]. This method determines the optimal threshold value based on the measured location and distribution of the intensity value of the foreground and background. This method selects the minimum criterion value by using the mean of their normal distributions and standard deviation (σ) by the estimation of location and distribution of pixels respectively.

The proposed method modifies the traditional MET method by including additional parameters and by selecting the maximum criterion value. This approach improved the result and produced well segmented MRI brain images. This paper organized as follows, proposed method explained in section 2. The results and discussion is given in section 3 and conclusion is given in section 4.

2 The Proposed Method

2.1 Kittler and Illingworth's Method

The algorithm of MET method is based on the Bayesian classification rule [13]. This method first compute the bi-model histogram of the gray level image $h(g)$

that is normally distributed. Next it estimates the priori probability P_i of gray level of histogram $h(g)$ and find the mean of total probability. It is the initial threshold (T) of given image and separates the image into foreground and background classes (i=1, 2...). Then it computes the mean of their normal distribution as μ_i and standard deviation σ_i by using the following equations.

$$P_i(T) = \sum_{g=a}^{b} h(g) \tag{1}$$

$$\mu_i(T) = \frac{1}{P_i(T)} \sum_{g=a}^{b} h(g) \tag{2}$$

$$\sigma_i^2(T) = \frac{1}{P_i(T)} \sum_{g=a}^{b} (g - \mu_i(T))^2 h(g) \tag{3}$$

where $a = \begin{cases} 0 & i = 0 \\ T+1 & i = 2 \end{cases}$ and $b = \begin{cases} T & i = 1 \\ n & i = 2 \end{cases}$

The criterion function is,

$$\begin{aligned} J(T) = 1 &+ 2[P_1(T) log \sigma_{bg}(T) + P_2(T) log \sigma_{fg}(T)] \\ &- 2[P_1(T) log P_1(T) + P_2(T) log P_2(T)] \end{aligned} \tag{4}$$

The threshold T_0 is calculated based on the minimization criterion of $J(T)$ and defined as:

$$T_0 = \ arg_{1 \le t \le n} min J(T) \tag{5}$$

2.2 Modified Kittler and Illingworth's Method

For the good segmentation of MRI brain images, foreground is the necessary region. The traditional MET method may not work properly in MRI brain images. The total variance of given image and new weight $(1 - P_i(T)/2)$ were included in the MET method. This weight can make the best threshold value of modified MET method. This total variance and weight are used to maximize the criterion value. The proposed method takes maximum criterion value of modified MET method. That maximum value selects the threshold value for proposed method. The modified parameters are defined as:

$$\sigma_i^2(T) = \frac{1}{P_i(T)} \sum_{i=0}^{n} (g - \mu_i(T))^2 h(g) \tag{6}$$

$$\mu_i(T) = \frac{1}{P_i(T)} \sum_{i=0}^{n} h(g) \tag{7}$$

where $\sigma_i^2(T)$ is represent the total variance and $\mu_i(T)$ represents the total mean of the given image. This proposed method, include new weight $(1 - P_i(T)/2)$ and variance that increase the value of minimum criterion of MET method. This method gives the satisfactory results of only for the MRI brain images, while compared to other threshold methods.

The modified criterion function is,

$$J(T) = 1 + 2\big[P_1(T)log(\sigma_{bg}(T) - \sigma_i^2(T)) + P_2(T)log(\sigma_{fg}(T) - \sigma_i^2(T))\big] - 2[P_1(T)logP_1(T) + P_2(T)logP_2(T)](1 - P_i(T)/2) \quad (8)$$

The maximum criterion of $J(T)$ is used to find threshold T_1,

$$T_1 = \ arg_{1 \le t \le n} maxJ(T) \quad (9)$$

3 Results and Discussion

The performance analysis carried by using some MRI brain images. MRI brain images were selected from The Whole Brain Atlas website maintained by Harvard Medical School, USA [14]. The segmented images are evaluated by using NU parameter. Ground truth information is not required for this measure and defined as,

$$NU = \frac{|F_{fg}|}{|F_{fg} + B_{bg}|} \frac{\sigma_{fg}^2}{\sigma^2} \quad (10)$$

where σ^2 represents the variance of whole image and σ_{fg}^2 represents the variance of foreground. A well segmented image will have nonuniformity close to 0. In worst case, $NU = 1$. The worst case corresponds to an image for which background and foreground are indistinguishable up to second order moments.

The proposed method is compared with other thresholding methods by using the final binary images of MRI brain images. In Fig.1, the MRI brain images are given in column 1, the results of Otsu's method are in column 2, the results of Ridler and Calvard's method given in column 3, the results of MET method given in column 4 and the results of modified MET method are given in column 5. Our proposed method gives good results for MRI brain images. The threshold and NU values were computed from the results of MRI brain images and shown in Table 1.

In Fig.2, the graph shows the quantitative representation of NU measure of different methods with line representation. Thus vertical axis represents NU value and horizontal axis represents slice number of MRI brain image volume. The NU value of well segmented image is close to 0. The NU value of traditional MET method is close to 1 and threshold value is minimum. Further, under segmented binary images are shown in column 4 in Fig.1. The global threshold segmentation method, Otsu, gives satisfactory results for MRI brain images while compared to Ridler and Calvards method and MET method. Our proposed method provides well segmented binary images and shown in column 5 in Fig.1. The NU value of

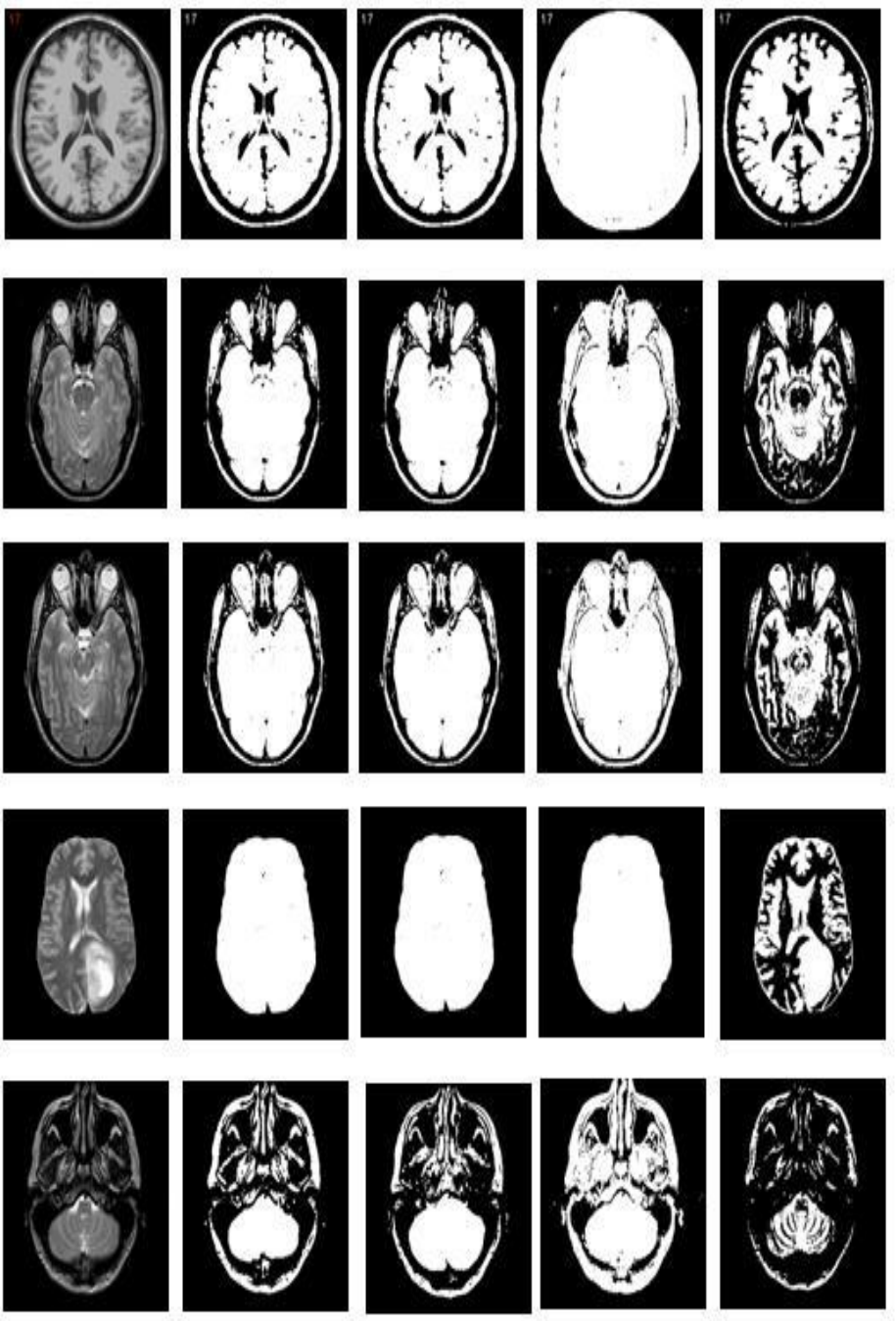


Fig. 1. Sample MRI brain images are in column 1, the results of Otsu's method are in column 2, the results of Ridler and Calvard's method are in column 3, the results of MET method are in column 4, and the results of Modified MET method are in column 5

Table 1. Region Non-Uniformity of MRI Brain Images

	Otsu		***Ridler and Calvard***		***MET***		***Modified MET***	
Images	T	NU	T	NU	T	NU	T	NU
BI-1	62	0.3940	59	0.4070	8	0.5848	108	0.1436
BI-2	62	0.4410	59	0.4551	8	0.6564	108	0.2020
BI-3	61	0.4317	59	0.4365	8	0.6495	108	0.2352
BI-4	59	0.3685	58	0.3635	8	0.6192	108	0.2319
BI-5	62	0.3146	60	0.3268	8	0.5855	108	0.2184
BI-6	58	0.3098	59	0.3120	8	0.5453	108	0.2380
BI-7	62	0.2604	61	0.2663	8	0.4647	108	0.2025
BI-8	62	0.2176	60	0.2247	8	0.3905	108	0.1428
BI-9	65	0.2013	63	0.2035	8	0.3360	108	0.1781
BI-10	65	0.2377	63	0.2390	8	0.3313	108	0.2119
BI-11	62	0.2331	61	0.2346	8	0.3068	108	0.2092
BI-12	62	0.2059	60	0.2068	8	0.2949	108	0.1991
BI-13	62	0.2053	61	0.2067	8	0.2938	108	0.2032
BI-14	62	0.2105	60	0.2118	8	0.2950	108	0.2041
BI-15	62	0.2259	60	0.2269	8	0.3085	108	0.2223
BI-16	62	0.2625	63	0.2645	8	0.3371	108	0.2273
BI-17	62	0.2693	64	0.2815	8	0.3495	108	0.2293
BI-18	58	0.2693	59	0.2721	8	0.3330	108	0.2126
BI-19	57	0.2115	55	0.2121	8	0.2958	108	0.1981
Average		**0.2773**		**0.2816**		**0.4198**		**0.2059**

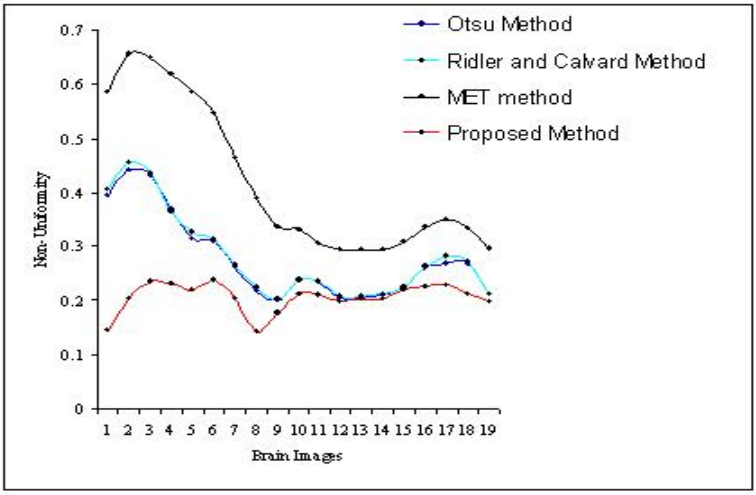


Fig. 2. Region Non-Uniformity Measurement Graph

proposed method is more close to 0 for MRI brain images, while compared to other methods. The threshold value of this method is maximum. The proposed method thus given the best results for MRI brain images.

4 Conclusion

This paper proposed a modified formula for popular thresholding method, MET and thus produced a robust thresholding technique for MRI brain image analysis. The resultant images produced by the proposed work were validated against the popular methods and traditional MET method. Experiments showed best results for modified proposed work than others. The modified method is well suited for segmenting the MRI brain images and analyzing its sub structures.

References

1. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson Education, Inc. Publication, New Delhi (2009)
2. Sonka, M., Hlavac, V., Boyal, R.: Digital Image Processing and Computer Vision. Cengage Learning (2008)
3. Kalaiselvi, T.: Brain Portion Extraction and Brain Abnormality Detection from Magnet Resonance Imaging of Human Head Scans. Pallavi Publications, Tamil Nadu (2011)
4. Al-amri, S.S., Kalyankar, N.V., Khamitkar, S.D.: Image Segmentation using Threshold Techniques. Journal of Computing 2 (2010)
5. Xue, J.H., Zhang, Y.J.: Ridler and Calvards, Kittler and Illingworths and Otsus Methods for Image Thresholding. Pattern Recognition Letters 33, 793–797 (2012)
6. Sahoo, P.K., Soltani, S., Wong, A.K.C.: Survey of Thresholding Techniques, Computer Vision. Graphics and Image Processing 41, 230–260 (1988)
7. Sauvola, J., Pietikainen, M.: Adaptive Document Image Binarization. Pattern Recognition 33(2), 225–236 (2000)
8. Otsu, N.: A Threshold Selection from Gray level Histograms. IEEE Transactions of systems, Man and Cybernetics (SMC) 9(1), 62–66 (1979)
9. Shaikh, S.H., Maiti, A.K., Chaki, N.: A New Image Binarization Method using Iterative Partitioning. Springer- Machine Vision and Applications (2012)
10. Nikolaos, N., Dimitris, V.: A binarization algorithm for historical manuscripts. In: 12th WSEAS International Conference on Communications, Heraklion, Greece, July 23–25, pp. 41–51 (2008)
11. Kalaiselvi, T., Nagaraja, P.: A Robust Thresholding Technique for Image Segmentation from General Gray Images. In: Proceedings of ICAMTCS–2013, pp. 183–188 (January 2013)
12. Sezgin, M., Sankur, B.: Survey Over Image thresholding Techniques and Quantitative Performance Evaluation. Journal of Electronic Imaging 13(1), 146–165 (2004)
13. Kittler, J., Illingworth, J.: Minimum Error Thresholding. Pattern Recognition 19, 41–47 (1986)
14. The Whole Brain Atlas (WBA), Department of Radiology and Neurology at Brigham and Womens Hospital, Harvard Medical School, Boston, USA

Histogram Based Split and Merge Framework for Shot Boundary Detection

D.S. Guru and Mahamad Suhil

Department of Studies in Computer Science, Manasagangothri, Mysore, India
dsg@compsci.uni-mysore.ac.in, mahamad45@yahoo.co.in

Abstract. In this paper, we propose a non-parametric approach for shot boundary detection in videos. The proposed method exploits the split and merge framework by the use of color histograms. Initially, every frame of the input video sequence undergoes color quantization and subsequently, the color histograms are computed for every quantized frame. The split and merge is driven by the fishers linear discriminant criterion function which results with a set of subsequences after several iterations which are assumed to be the shots present in the given video. The proposed method is experimentally tested on video samples from TrecVid 2002 dataset and YouTube online database. We have obtained overall accuracy of 85.5% Precision, 87.1% Recall and 86.1% F-measure for the dataset used. A comparative study of the proposed approach with the contemporary research works is also carried out.

Keywords: color quantization, color histograms, split and merge, fishers linear discriminant analysis, shot boundary detection.

1 Introduction

From the past two decades, due to the rapid development of digital storage technology and available bandwidth, the activities such as storing, sharing and searching multimedia data over the internet have become indispensable components of our life. Among all the multimedia data, video is frequently used since it preserves both visual and temporal behaviors of objects present in a scene. But, searching for a video of a particular interest is highly difficult and time consuming due to the size of the multimedia database available on the web. To manage such a huge multimedia database, for the last two decades, there have been a couple of attempts towards development of automated content based video indexing and retrieval (CBVR) systems [1-6].

Shot boundary is the location where the transition from one shot to the other subsequent shot takes place in a video [7]. Among all the various steps involved in CBVR system, shot boundary detection stage is dealt very carefully and it is the first step in the CBVR system and its efficiency is very much necessary for further activities such as, key frame extraction, video indexing, dimensionality reduction and

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 180–191, 2013.
© Springer International Publishing Switzerland 2013

representation of the video[6]. Abrupt and Gradual transitions are the two types of shot transitions available in videos [9].

In literature, we can find a couple of good attempts to solve the problem of shot boundary detection based on various representation techniques. Some of the major approaches are: histograms based approaches[10-11] where, the similarities and continuity of the frames in a sequence are measured with the help of differences of histograms to arrive at the possible locations of maximum discontinuities, block based approaches [12] where each video frame is studied at the block level to extract local features and matched with the corresponding blocks of the subsequent frames for the identification of shot change, model based approaches [13-14] where a model is trained to identify the possible shots, cluster based approaches [15-17] where frame sequence is clustered into several clusters and every cluster is checked for the possibility of being a shot, non-parametric approaches[18] where shot boundaries are detected without consuming any parameters such as a threshold, compressed domain approaches[19-21] where the video is processed in its compressed domain itself so that, the time of decompression is avoided, fusion based approaches [22-24] where, several approaches are fused with different combinations to make use of the advantages of various popular techniques and so on. Various features such as color, texture, shape, sketch, SIFT, motion vectors, edges in spatial as well as in transformed domains such as Fourier, cosine wavelets, Eigen values, etc., are used majorly with different combinations of the same in many popular approaches.

Regardless of the extensive investigations made and copious techniques proposed, shot boundary detection is still an active area of research with many challenges [25-27]. It is because of the fact that the researchers are unsuccessful in arriving at a universal and robust model for shot boundary detection which can be an ideal solution for the problem of shot boundary detection in videos of any modality with any amount of complexities being present. With this insight, we have attempted to solve the problem of shot boundary detection with split and merge framework. In our method, we have used a memory efficient color histogram representation to represent the video frames in reduced dimensions. Followed by this, we have exploited the split and merge framework introduced in [28] for automatic shot detection from videos. The proposed method is experimentally validated on videos of Trecvid dataset and some of the videos downloaded randomly from internet. A qualitative comparative analysis of the proposed approach with the contemporary research works is also carried out.

The rest of the paper is organized as follows: Section 2 presents the proposed shot boundary detection method. Experimental results and analysis are discussed in section 3. The conclusion and future work are given in section 4.

2 Proposed Method

In this section, we propose our method for shot boundary detection using the split and merge framework by using color histogram features as video frame representatives.

Given a video, on an average there will be 25 frames available per second leading to a very huge sequence. Handling and processing such a long sequence of frames with very huge dimension is highly computing time and space demanding task. Hence, the dimensionality reduction needs to be done by extracting only the discriminating features.

Fig. 1 shows the major steps involved in the proposed method. Given a video, we represent every video frame using color histogram features. The sequence of video frames is then treated as a sequence of feature vectors which are fed to the split and merge framework to get split into several subsequences so that each subsequence is complete in its own sense and capable of being an individual unit in the given video. These subsequences will later be proved as the shots present in the given video. So, after completion of split and merge process using the sequence of video frame representatives, we can easily identify the boundaries of each and every shot of that video.

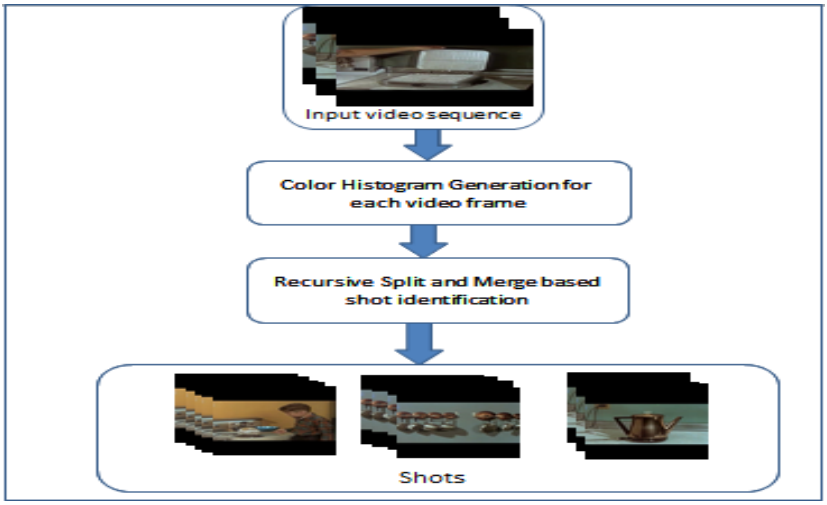


Fig. 1. Block Diagram of the Proposed Method

2.1 Color Histogram for Video Frame Representation

In [28], the authors reduce the dimensionality of the input video frame sequence by using Haralick texture features and spectral clustering techniques. Initially every frame is divided into 32X32 blocks and Haralick texture features are extracted from each block. After this, the extracted set of texture features of all blocks of a frame are clustered together using spectral clustering so that, if any natural region which was earlier divided because of block division can now be combined upon comparing the texture features of corresponding blocks. With this a video frame with size mXn is represented with the help of only k feature vectors corresponding to k clusters of the

frame and the dimension of the extracted feature vectors was very less when compared to that of the original frame.

The major disadvantage of the work proposed in [28] was the amount of time required to process each video frame since they process every video frame in various levels to get its dimensionality reduced. In our present method, we try to extract features out of each video frame through the color histogram which can be processed in less time.

Color is one of the most significant properties of images. Images and videos can be efficiently indexed through the use of color information present in them. Color information can be extracted easily when compared to most of the image features and color is upto certain level invariant to transformations like translations, rotations, mirroring and scaling.

We are motivated by the work of Mas et al., in [29] where, a color histogram based shot boundary detection method is proposed. The method was based on the image bit plane slicing philosophy [30]. When a gray scale digital image with 8-bit per pixel is given, all 8 bits may not contribute equally to the appearance of the image but, the higher order bits affect significantly when compared to the lower order bits. Hence, eliminating a certain number least significant bits from each pixel can reduce the memory requirements of the image which is also a method of quantization used in image compression. The authors of [29] have used the same philosophy to compress the video frames. Initially, every video frame in RGB color space represented with 24 bits/ pixel is considered and 4 least significant bits from each of the R, G and B color space are removed. After elimination, the frame needs only 12 bits/pixel for representation. By using the quantized video frame, the authors have proposed to create a color histogram. After having efficiently represented every video frame with only 50 percentage of its original size, a color histogram for the reduced 12 bit equivalent frame is created. Since there are only 212(=4096) different possible values for each pixel in the quantized video frame, the number of histogram bins required for the creation of histogram is only 4096, which is very much less compared to the original number of bins.

We make use of the color histograms generated in the same way as video frame representatives in our work. The Fig. 2 shows a few sample images in RGB color space with 24 bits/Pixel and their corresponding 12 bits/Pixel representation. It is evident from the images with 12 bits/Pixel that we can visualize all important components in the image even after eliminating 12 least significant bits.

With this, given a video frame with any dimension, we can generate a color histogram based representative for it with only 4096 values corresponding to 4096 different color bins. So, a video with dimension N*(m*n*3), where N is the number frames, with each frames composed of three mXn matrices each of which is correspond to a particular color band in RGB color space, can be reduced to N feature vectors with 4096 values.

(a) (b)

Fig. 2. Sample Video Frames. (a). With 24 bits/ pixel. (b). with 12 bits / pixel

2.2 Shot Boundary Detection Using Split and Merge Framework

We now apply the process of split and merge to the input video sequence with N frames which is represented with N color histograms as explained in the section 2.1.

We first present the overview of split and merge framework introduced in [28]. In this framework, the authors have approached the shot boundary detection problem in a new way using split and merge concept. Split and merge is a well-known algorithmic strategy used to solve any complex problem, where the problem is subdivided into smaller sub problems through successive splitting and merging until each sub problem becomes atomic, that it can be solved without any difficulty. In this framework, the same analogy is used to arrive at the shots from a given video. The larger sequence of video frames is subdivided into smaller subsequences repeatedly

through split and merge using their representatives until the video is divided into many smaller subsequences each of which is dissimilar from the other and the fishers criterion function between every two adjacent subsequences is high. The Fishers Linear Discriminant analysis is chosen in the framework because of the fact that FLD is the best known technique available in the literature for finding out the optimal projection axis between two classes projected on to a space [31-32]. If the data points are well separated in the projected space, then the value of the criterion function for those two classes will be maximum. In case of video frames, if we can project the samples from each of the two subsequences that is being tried for split on to some lower dimensional space and if the optimal projection axis is found between those two classes of data points, then the fishers criterion between those two subsequences will be maximum. Finally these smaller subsequences are declared to be the shots present in the given video.

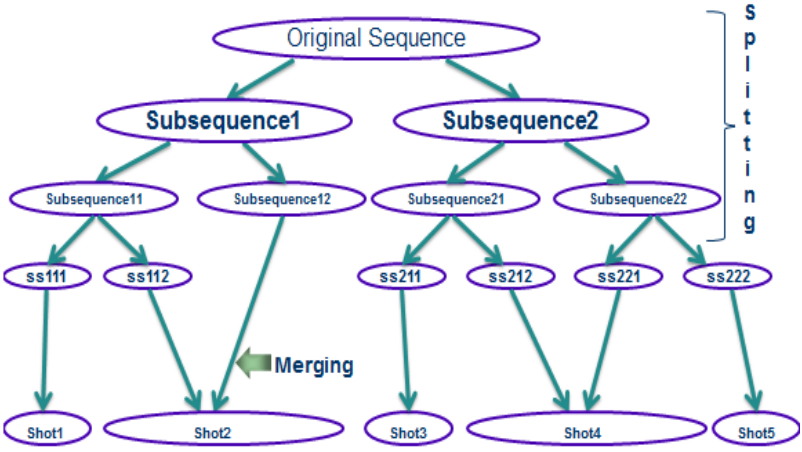


Fig. 3. Split and Merge Framework

A video sequence, represented with the help of histogram representatives is given to the split and merge framework for recursive split and merge. Initially, the entire sequence of frames is allowed to get split into exactly two subsequences. Then, that division is validated by dividing each of the subsequence again into exactly two and comparing the value of the fishers criterion function between each pair of adjacent subsequences with that of the previous value obtained in the first split. If the values between current pair of adjacent subsequences are higher compared to that of the previous level then only the process is allowed to continue further with each smaller subsequence. Otherwise, the corresponding split is rolled back so that, the two smaller subsequences after splitting are combined back to form a single subsequence. Along with splitting, in each iteration we also check whether any two adjacent subsequences can be combined into single with the help of the same fishers criterion. That is, if the

value of the fishers criterion function gets increased or remains unchanged when any two adjacent subsequences are combined into single then, they are allowed to merge into single subsequence. The reason which allows us to split a sequence into two subsequences or merge any two smaller subsequences into one is as follows: Whenever there is a shot boundary present in a given video sequence, if we split it into two and project it onto a lower dimensional space, we can find them projected as two different clusters because of high between class scatter and less within class scatter. After several split and merge actions repeated up to a set of subsequences with only a single frame in each in the worst scenario, the resultant small subsequences of the given video are then can be proved as the shots present in the videos.

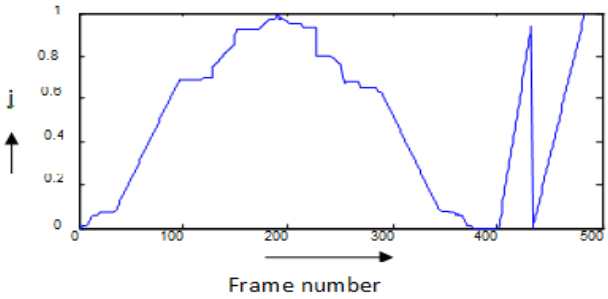


Fig. 4. Variations in J value with respect to frame sequence

The Fig. 4 shows the curve of value of fishers criterion function J with respect to a sample video frame sequence with number of frames equal to 500. A value at the location (x, y) on the curve in the figure represents the value of J when the video frame sequence with n frames is projected as two classes with first class consisting of a subsequence of frames from frame 1 to frame x and the second consisting of subsequence of frames from frame x+1 to frame n. Where, x represents the frame number and y represents the value of J. The locations, where the J values are maximum represent that, the corresponding two subsequences are well projected in the space.

3 Experimentation and Analysis

To evaluate the suitability of the proposed approach, we have experimented on seven different video samples. Four video samples are taken from the TrecVid 2002 dataset (available at http://www.open-video.org/) and three video samples are downloaded from YouTube in which one is news other is cricket and one more is an animated movie video. Manually identified shots present in each of the testing video sequence are considered as ground truth. The dataset and ground truth details are given in the Table 1.

Table 1. Dataset and Ground-truth

Dataset Source	Serial Number	Caption	Number of Frames Considered	Length in Secs	No. of Transitions
1. TrecVid 2002	1	17 Days: The Story of Newspaper History in making (1945)	4000	137	15
	2	1955 Chevrolet Screen Ads	3900	135	12
	3	6 1/2 Magic Hours	4800	166	16
	4	According to Plan: The Story of Modern Sidewalls for the Homes of America	5200	180	17
2. Internet Videos	5	News	2760	120	30
	6	Cricket	2000	80	16
	7	Animated Movie	3000	131	34

The results obtained by the proposed approach for test dataset is given in the Table 2. We make use of the following most popular evaluation measures to evaluate the results of the proposed method,

$$\text{Precision (P)} = \frac{D}{D + FA'}, \quad \text{Recall} = \frac{D}{D+MD'}, \quad \text{F-measure (F)} = \frac{2 * P * R}{P + R}$$

Where D is the number of shots correctly detected, MD is number of shots missed (MD), and FA is the number of false alarms.

We can notice from the Table 2 that, on an average the overall performance of the proposed method is really good. Among all seven videos considered, a very high value of F-measure is achieved for the fourth video which is from the TrecVid dataset as there is only one shot missed with number of false alarms being one. This is because, the video is free from object motion and also it has a less number of gradual transitions in it. The method has very less performance compared to all others in the case of video 6 a cricket video. This is because of the fact that the cricket videos are created with a very high amount of editing effects like score board or match summary being displayed suddenly during the broadcast of video, zooming effects, high illumination effects because of outdoor environment, object motion, camera motion and usage of multiple cameras.

Table 2. Experimental results of the proposed method

Serial Number	D	MD	FA	PRECISION (P)	RECALL (R)	F-MEASURE (F)
1	12	3	2	0.857	0.800	0.828
2	11	1	3	0.786	0.917	0.846
3	13	3	2	0.867	0.813	0.839
4	16	1	1	0.941	0.941	0.941
5	26	4	3	0.897	0.867	0.881
6	14	2	4	0.778	0.875	0.824
7	30	4	5	0.857	0.882	0.870
			Average	0.855	0.871	0.861

Table 3. Qualitative Comparative Analysis

Method	Video Representation	Parametric	Parameters Used	Remarks
[33]	Gray level frames are divided into 16X16 macro blocks and block motions are estimated and a plot of the differences between two successive motion intensities is used	yes	Number of blocks per frame, Number of adjacent blocks to be compared for motion estimation, T: Matching error threshold between macro blocks	Block motion estimation requires comparison with 5 neighboring blocks. Matching error threshold requires training.
[34]	Histogram of each of the R G B color bands in RGB space	yes	d: Number of bins in histogram, n: Number of blocks, r: Range of neighboring frames involved in the SBD, r1: range of frames involved in the computation of continuity feature, Tp: Threshold for similarity between histograms of two frames, T: range of frames involved for the classification of SBDs, T0: threshold for monochrome value, f: neighbored range, Ts: threshold for similarity value, Tα: threshold for distance, Tm: threshold for monochrome values.	Even though a high accuracy is assured, due to the huge number of parameters make it a complex and fixing those parameters needs rigorous training. They have used two SVM's and the complexity of each is O(n3) where, n is the training set size.
[29]	Color Histogram from a color quantized frame	yes	W: window size used for convolution of difference of histograms signal, A threshold to detect cuts, α: weight factor with range [0, 1], size of the structuring element for signal convolution and morphological operations,	Through the use of color histograms followed by color quantization has made the method memory efficient. But, use of certain parameters and processing the signal at various stages has made it a slightly complex approach.
Proposed Approach	Color Histogram from a color quantized frame	No	Nil	In addition to being memory efficient, the approach is non- parametric with the use of split and merge framework.

To demonstrate the superiority of the proposed method, a qualitative comparative analysis with the contemporary research works [29, 33-34] has been done in Table 3. The work [33] is considered because it uses block motion technique to estimate the motion vector for each video frame which is proven to be effective with high accuracy. The work [34] is considered because, it uses a variant of color histogram creation process to extract the features from video frames. Another approach which resembles our approach in terms of features used is the work presented in [29]. In this work the similar histogram features are used to represent a video. But, they use histogram differences curve to locate shot boundaries with some post processing applied using convolution, derivatives and morphological operators. In our model, we treat histograms as representatives of video frames and apply split and merge process to segment the video into its constituent shots. It can be noted from the comparative analysis that, the proposed approach is very well suited for the real time applications as it is free from parameters making it a non-parametric approach in addition to being time efficient.

4 Conclusion

In this work, we have exploited the split and merge framework for shot boundary detection. The theme of this framework is to view the video as contiguous groups of frames where frames within a group are more similar and preserve temporal continuity when compared to the frames in different groups. And these groups are nothing but the shots present in the video. The process of shot boundary detection is just the identification of the place where two contiguous groups of frames intersect. Here, the video is given to the split and merge framework by using color histogram representatives through the application of color quantization on to each video frame. Experiments are conducted on various video samples from Trecvid 2002 dataset and YouTube online database. The proposed method is evaluated using Precision, Recall and F-measures and we have achieved 85.5% of Precision, 87.1% of Recall and 86.1% of F-measure for the test data set considered. The qualitative comparative analysis is conducted with the contemporary research work and shown that the proposed method is better in various aspects.

References

1. Idris, F., Panchanathan, S.: Review of image and video indexing techniques. J. Vis. Commun. Image Represent. 8(2), 146–166 (1997)
2. Brunelli, R., Mich, O., Modena, C.M.: A survey on the automatic indexing of video data. J. Vis. Commun. Image Represent. 10, 78–112 (1999)
3. Koprinska, I., Carrato, S.: Temporal video segmentation: a survey. Signal Processing: Image Communication 16(5), 477–500 (2001)
4. Lefevre, S., Holler, J., Vincent, N.: A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. Real-Time Imaging 9(1), 73–98 (2003)

5. Patel, B.V., Shah, B.B.: Content based video retrieval systems. Int. J. Ubi Comp. 3(2), 13–30 (2012)
6. Kanagavalli, R., Duraiswamy, K.: A study on techniques used in digital video for shot segmentation and content based video retrieval. European Journal of Scientific Research 69(3), 370–380 (2012)
7. Mittal, A., Cheong, L., Sing, L.: Robust identification of gradual shot-transition types. In: Proceedings of 2002 International Conference on Image Processing, vol. 2, pp. 413–416 (2002)
8. Patel, N.V., Sethi, I.K.: Video shot detection and characterization for video databases. Pattern Recognition 30, 583–592 (1997)
9. Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., Zhang, B.: A formal study of shot boundary detection. IEEE Trans. on Circuits and Systems for Video Technology 17(2), 168–186 (2007)
10. Zhang, C., Wang, W.A.: Robust and efficient shot boundary detection approach based on fisher criterion. In: Proceedings of the 20th ACM International Conference on Multimedia (MM 2012), pp. 701–704. ACM, New York (2012)
11. Onur, K., Ugur, G., Ozgur, U.: Fuzzy color histogram-based video segmentation. Computer Vision and Image Understanding 114(1), 125–134 (2010)
12. Abdelati, M.A., Ben, A.A., Mtibaa, A.: Video shot boundary detection using motion activity descriptor. J. Telecommun. 2(1), 54–59 (2010)
13. Chen, W., Zhang, Y.: Parametric model for video content analysis. Pattern Recogn. Lett. 29(3), 181–191 (2008)
14. Massimiliano, A., Chianese, A., Moscato, V., Sansone, L.: A formal model for video shot segmentation and its application via animate vision. Multimedia Tools Appl. 24(3), 253–272 (2004)
15. Damnjanovic, U., Izquierdo, E., Grzegorzek, M.: Shot boundary detection using spectral clustering. In: 15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, pp. 1779–1783 (2007)
16. Wang, P., Liu, Z., Yang, S.: Investigation on unsupervised clustering algorithms for video shot categorization. Journal of Soft Comput. 11(4), 355–360 (2006)
17. Yuchou, C., Lee, D.J., Yi, H., James, A.: Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor. J. Image Video Proc. 1, 1–10 (2008)
18. Manjunath, S., Guru, D.S., Suraj, M.G., Harish, B.S.: A non-parametric shot boundary detection: an Eigen gap based approach. In: Proceedings of Fourth Annual ACM Bangalore Conference, vol. 1, pp. 1030–1036
19. Wang, H., Divakaran, A., Vetro, A., Chang, S.F., Sun, H.: Survey of compressed-domain features used in audio-visual indexing and analysis. J. Visual. Commun. Image Represent. 14, 150–183 (2003)
20. Bruyne, S.D., Deursen, D.V., Cock, J.D., Neve, W.D., Lambert, P., Walle, R.V.D.: A compressed-domain approach for shot boundary detection on H.264/AVC bit streams. Signal Processing: Image Communication 23, 473–489 (2008)
21. Chen, J., Ren, J., Jiang, J.: Modelling of content-aware indicators for effective determination of shot boundaries in compressed MPEG videos. Multimedia Tools Appl. 54(2), 219–239 (2011)
22. Jacobs, A., Miene, A., Ioannidis, G.T., Herzog, O.: Automatic shot boundary detection combining color, edge, and motion features of adjacent frames (2004), http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/ubremen.pdf

23. Chang, Y., Lee, D.J., Hong, Y., Archibald, J.: Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor. J. Image Video Process. 9, 10 (2008)
24. Philips, M., Wolf, W.: A multi-attribute shot segmentation algorithm for video programs. Telecommunication Systems 9(3-4), 393–402 (1998)
25. Boreczky, J.S., Rowe, L.A.: Comparison of video shot boundary detection techniques. J. Electron Imaging 5(2), 122–128 (1996)
26. Alan, F.S., Palu, O., Aiden, R.D.: Video shot boundary detection: Seven years of TRECVid activity. Comput. Vis. Image Und. 114(4), 411–418 (2010)
27. Mishra, R., Singhai, S.: A review on different methods of video shot boundary detection. International Journal of Management IT and Engineering 2(9), 46–57 (2012)
28. Guru, D.S., Suhil, M., Lolika, P.: A novel approach for shot boundary detection in videos. In: Multimedia processing, communication and computing applications. LNEE, vol. 213, pp. 209–220. Springer (2013)
29. Mas, J., Fernandez, G.: Video shot boundary detection using color histogram (2003), http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/ramonlull.paper.pdf
30. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. PHI Learning Private Limited, New Delhi-110001 (2008)
31. Max Welling.:Fisher linear discriminant analysis. Max welling's classnotes in machine learning. 16(7), 817–830, http://www.ics.uci.edu/~welling/classnotes/classnotes.html
32. Nagabhushana, P., Guru, D.S., Shekara, B.H. (2D)2 FLD: An efficient approach for appearance based object recognition. Neurocomputing. 69, 934–940 (2006)
33. Atmel, A.M., Abdessalem, B.A., Abdellatif, M.: Video shot boundary detection using motion activity descriptor. Journal of Telecommunications. 2(1), 54–59 (2010)
34. Zhang, C., Wang, W.: A robust and efficient shot boundary detection approach based on fisher criterion. In: Proceedings of 20th ACM International Conference on Multimedia, vol. 5, pp. 701–704 (2012) ISBN: 978-1-4503-1089-5, doi:10.1145/2393347.2396291

Boundary Detection of Objects in Digital Images Using Bit-Planes and Threshold Modified Canny Method

P. Shanmugavadivu and Ashish Kumar

Department of Computer Science and Applications,
Gandhigram Rural Institute - Deemed University, Gandhigram - 624 302,
Tamil Nadu, India
{psvadivu67,ashishgru}@gmail.com

Abstract. Two novel Canny-based boundary detection techniques are presented in this paper. Canny edge detection has gained popularity over the period due to its potential in edge detection. However, the edges detected by Canny are highly superfluous to extract the boundary of the objects in an image. The Modified Canny methods address this issue by modifying the parameter of Canny. The first method namely Threshold Modified Canny (MC-T) uses the Mean of the input image as threshold. MC-T is found to produce the boundaries even on the high-contrast images. The Second method, Bit-planes and Threshold Modified Canny (MC-BT) performs edge detection on the three intensity significant bit-planes using Mean of the input image as Threshold. This technique has also produced promising results in detecting the image boundary. The second method as it works only on three bit planes information of the input image, it reduces insignificant details and yields significant object boundaries. The result of the two proposed techniques, suitably finds place in object recognition, pattern recognition / matching etc. where boundary detection is an important component. These approaches are much promising in terms of clear boundary detection of an object, as boundary detection by conventional methods is very time consuming.

Keywords: Edge Detection, Bit-planes, Canny Algorithm, Modified Threshold.

1 Introduction

Edge detection is the most essential operation in image analysis. An edge is a manifestation of discontinuity of local characteristics (gray mutation, color mutation, mutation texture, etc), of an image. It defines the boundary between objects and their background. Edge detection is a vital preprocessing technique for image segmentation, object recognition pattern matching and computer vision. There are several algorithms developed for edge detection such as classical methods (Sobel, Prewitt, Kirsh etc) which are simple but sensitive to noise, Zero-Crossing methods (Laplacian, Second directional derivative) which detect

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 192–200, 2013.
© Springer International Publishing Switzerland 2013

edges in their orientation with fixed characteristics in all directions though it is also sensitive to noise, Laplacian of Gaussian (Marr-Hildreth) that finds correct places of edges and tests wider area around the pixels, malfunctions at the non-linear features such as corners, curves etc., Gaussian (Canny, Shen-Castan) which uses the probability for finding error rate, Localization and response, improving signal to noise ratio, exhibits better detection in noisy environment, it suffers from computational and time complexity and false zero-crossing [1, 4]. Commonly it is observed that Canny produces better results with the trade-off of time and space complexity [2, 4]. Usually most of the significant details of an object are found in most significant bit planes of any digital image. In this paper, a new approach is presented to detect the object boundary. This exploits those details and thereby produces interesting results. Section 2 discusses the principle of popular edge detection methods and their applications whereas Section 3 presents the methodology of the proposed MC-T and MC-BT. Section 4 presents the experimental results and discussion and the conclusions are drawn in Section 5.

2 Conventional Methods of Edge Detection

Typically an edge in an image is a boundary between an object and its background. Mathematically an edge can be represented by First and Second Order differential equations. The first-order derivative (i.e. Gradient) of a 2-D function f(x, y) is given by:

$$\nabla f = \begin{bmatrix} Gx \\ Gy \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{1}$$

where Gx and Gy are the gradients in the x and y coordinates, respectively. The magnitude of the vector is given by:

$$mag(\nabla f) = \sqrt{(Gx) + (Gy)} = \sqrt{(\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2} \tag{2}$$

Generally, the variance of the gray level is calculated with one of these edge detection operators or kernel operators. The slopes in the x and y directions are combined to give the total value of the edge strength. The edge detection operator fixes a kernel centered on a pixel chosen. If the value of this kernel area is above a given threshold, then the center pixel is classified as an edge [1]. The popular edge-detection methods classified in following categories, based on differential operators:
Gradient Detector (First order derivatives)
Zero crossing (Second order derivatives)
Laplacian of Gaussian (LoG)
Gaussian Edge Detector
Colored Edge Detector

Gradient edge detector uses first order derivatives as in Sobel (1970), Prewitt (1970), Kirsch (1971), Robinson (1977), Frei-Chen (1977), Deatsch and Fram (1978), Nevatia and Babu (1980), Ikonomopoulos (1982), Davies (1986), Kitchen and Malin (1989), Hancock and Kittler (1990), Woodhall and Linquist (1998) and Young-won and Udpa (1999) [5,7]. Zero-Crossing uses second order derivatives and Laplacian operators. Laplacian of Gaussian (LoG) was invented by Marr and Hildreth (1980) which combines Gaussian with Laplacian filter. LoG is seldom used in machine vi-sion. Berzins (1984), Shah, Sood and Jain (1986), Huertas and Medioni (1986) have worked on LoG[2]. Gaussian edge detectors is used by Canny and ISEF (Shen-Castan), it works symmetrically along the edges and smoothens the image to reduce noise [2, 8, 9].

3 Proposed Method

In general, any digital image is processed as a whole to detect the edges and it produces lot of unwanted edges which may mislead the object recognition algorithms applied on the edge detected images. A simple and effective approach for boundary edge detection is devised, with the objectives to reduce unwanted edges. This technique detects boundary edges of objects which ultimately reduce the preprocessing time substantially and thus highly suits for real-time application. All digital images are formed with the combination of bit planes. For example, binary, gray, RGB images are formed with 1, 8, 24 (8 bit for each red, blue and green) bit planes respectively. Usually only a few Most Significant Bit-planes contain majority of edge information. So those bit-planes can be subjected to edge detection. Proposed technique of edge detection works with two algorithms one for threshold selection and another for bit-planes selection. The algorithmic description is given below.

1. Algorithm for MC-T

Input : Image (I)
Output : Required Threshold (T) and Object Boundary (OIT) of original image
1. Read image I
2. Convert I into gray image if not in gray
3. Compute histogram of image I with Ci (count) for each intensity 'i', where $0 \leq i < 256$
4. calculate threshold modified value using the formula

$$T = \frac{\sum_{i=0}^{255} C_i * i}{(row * column)} \tag{3}$$

where Ci = Number of counts for intensity 'i'
5. Apply Canny on I with threshold (T)
6. Display Modified Canny with 'T' (OIT)

For MC-T we accept an original digital image and convert it into gray image if it is not gray image. Compute histogram of image I with Ci (count) for each intensity 'i', where $0 \leq i < 256$ and calculate threshold modified value (T) using the formula in step (4). Now apply Canny on original image (I) with T. Display the results of MC-T that produces convincing results for boundary detection. As an incremental modification, MC-BT was developed.

2. Algorithm for MC-BT

Input : Image (I), Threshold modified (T)
Output : Object Boundary (OIBT) of bit-planes image

1. Read image (I)
2. Convert I into gray image if not in gray
3. Compute Histogram of image (I) with 8 bins [level 0, level 1 level 7]
4. Select top three bins whose count values is the maximum
5. Consider those bit-planes whose bins are selected in step (4)
6. Combine the selected bit-planes as B = (B1, B2,& B3) for edge detection and convert it into gray image
7. Apply Canny on B with threshold (T)
8. Display Modified Canny on bit-planes image (B) with 'T'(OIBT)

For bit-planes selection, accept original image (I) and modified threshold (T) and convert (I) into gray image if not in gray. Calculate histogram with eight bins and select top three bins which have high-count values. Combine selected bit-planes based on step (5). Apply Canny on gray image obtained from 3-bit-planes B = (B1, B2, & B3) of (I) with the computed threshold modified (T). Display the results of Modified Canny on bit-plane's gray image (B) with MC-BT.

4 Results and Discussion

The proposed algorithms are implemented using Matlab-7. Images of different intensity distribution with respect to foreground and background were chosen from the internet. Fig. 1(a) is showing one fighter plane flying over mountain range.

Similarly in Fig. 2(a) one fighter is flying over the coastal region having water and sand background and Fig. 3(a) shows another plane with sky. These input images Fig. 1(a), Fig. 2(a), Fig. 3(a), Fig. 4(a) were processed by Canny, MC-T and MC-BT and the respective results are shown in Fig. 1(b), Fig. 2(b), Fig. 3(b), Fig. 4(b) for Canny, Fig. 1(c), Fig. 2(c), Fig. 3(c), Fig. 4(c) for MC-T, Fig. 1(d), Fig. 2(d), Fig. 3(d), Fig. 4(d) for MC-BT. It is evident from the results of Fig.1, 2, 3 and 4 that the proposed algorithms bring out the edges of the object of interest in an image more precisely than the Standard Canny

method. The result of both MC-T and MC-BT are free of unwanted edge details, and can be used by object recognition algorithms applied on the edge detected images. Thus, the newly devised techniques presented in this paper are more suitable to be integrated into the edge-based object recognition methods. The MC-T and MC-BT techniques clearly define the outline of the objects using which, the shape, size and orientation can be found, which are the essential features for object recognition, pattern recognition, pattern matching etc. However, the minor drawback of these techniques is that, when the intensity difference between the object and background is minimum, it is unable to highlight the edges which get merged with the background Fig. 4(c) and Fig. 4(d). This problem can be addressed, by taking the intensity details of one or more bit-planes in addition to three selected bit-planes. Another solution to this issue is to enhance the low contrast images using simple histogram equalization and then proposed methods can be applied on the enhanced images.

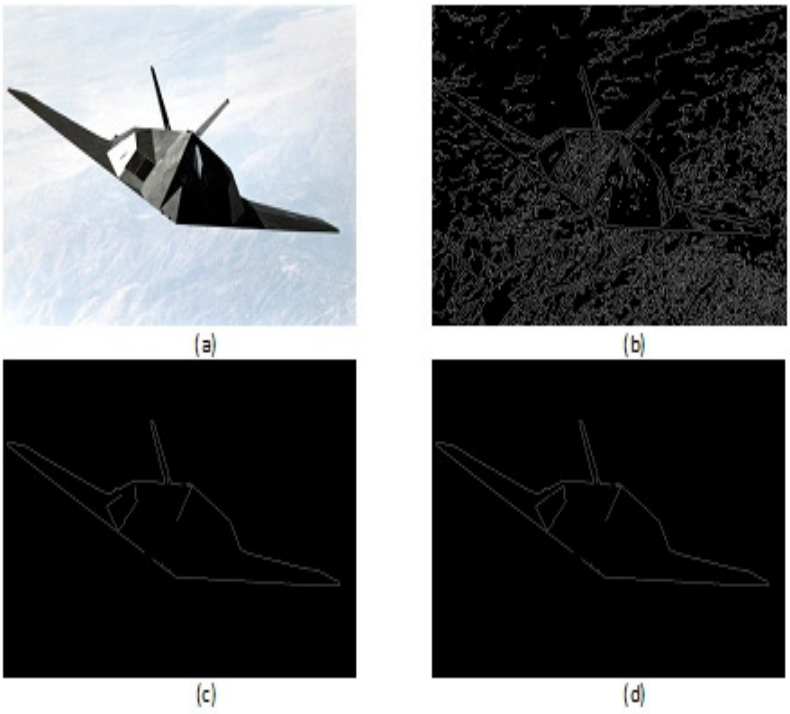


Fig. 1. (a) : Original
(b): Canny Method (c): Modified Canny - T (d): Modified Canny - BT

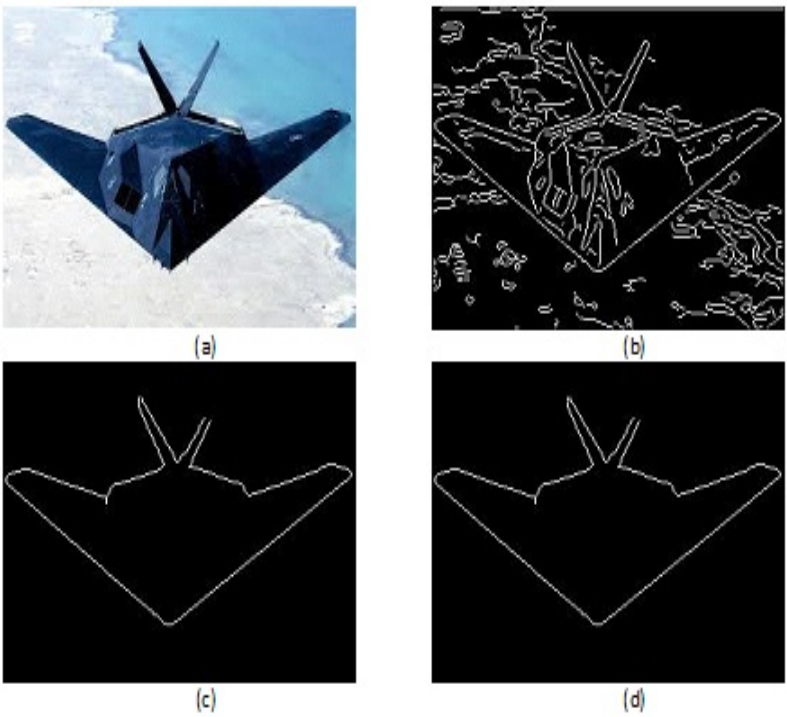


Fig. 2. (a) : Original
(b): Canny Method (c): Modified Canny - T (d): Modified Canny - BT

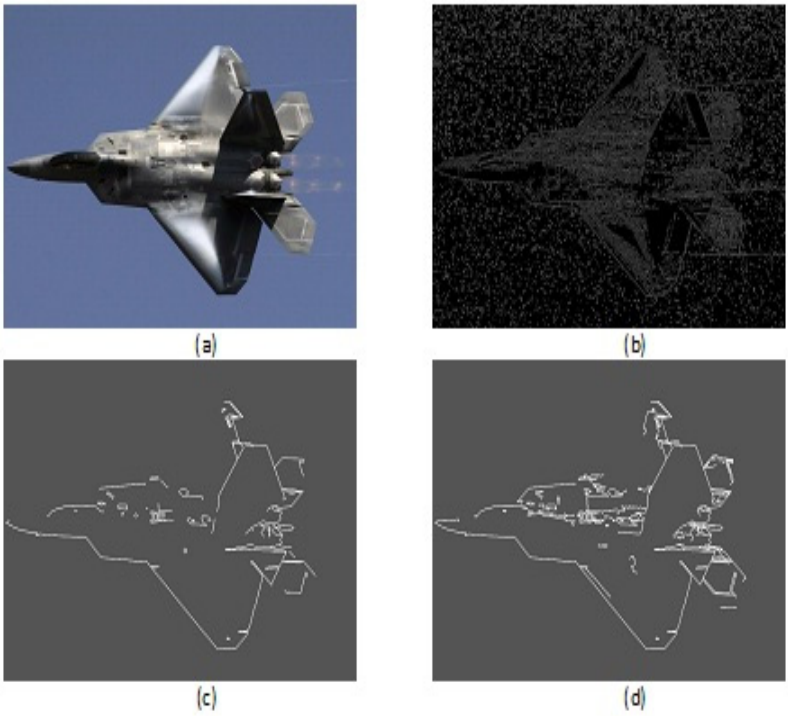


Fig. 3. (a) : Original
(b): Canny Method (c): Modified Canny - T (d): Modified Canny - BT

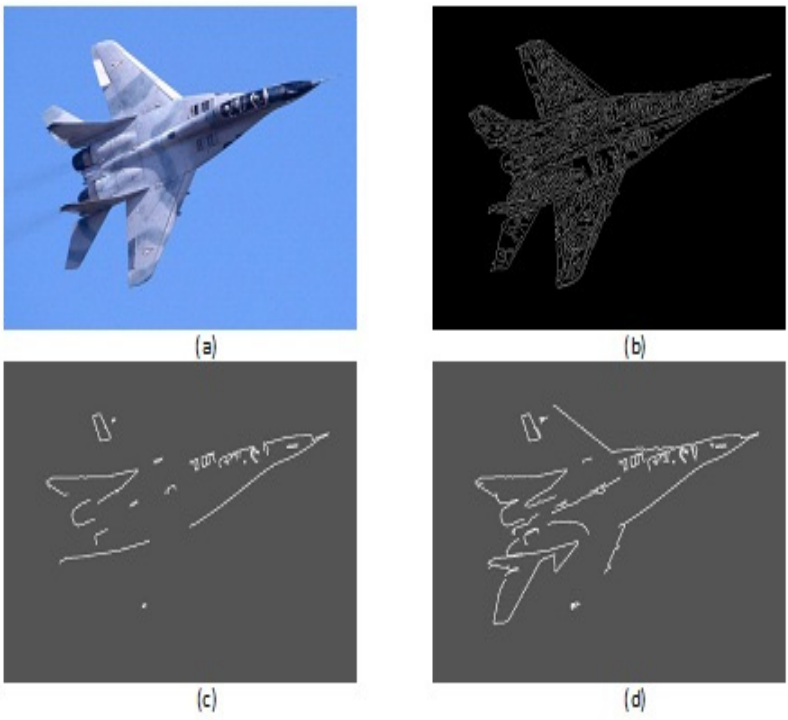


Fig. 4. (a) : Original
(b): Canny Method (c): Modified Canny - T (d): Modified Canny - BT

5 Conclusions

In this paper two novel modified Canny-based edge detection techniques are presented namely, Threshold Modified Canny (MC-T) and Bit-plane and Threshold Modified Canny (MC-BT), which are proved to be effective in detecting the boundaries of the objects. These methods perfectly work on high contrast images, when compared to low contrast images. However, these devised algorithms are proved to be highly suitable for precise object detection / pattern recognition than canny method, as the former one suppresses the intrinsic edges which may highly mislead. The major advantage of the proposed techniques MC-T and MC-BT are clear and precise detection of boundary of objects.

References

1. Rafael Gonzalez, C., Richard Woods, E.: Digital Image Processing, 2nd edn. Pearson Education, New Delhi (2002)
2. Bovik, A.: Handbook of Image and Video Processing, 2nd edn. Academic Press (2005)
3. Bao, P., Zhang, L., Wu, X.: Canny Edge Detection Enhancement by Scale Multiplication. IEEE Trans. Pattern Anal. Mach. Intell. PAMI–27(9), 1485–1490 (2005)
4. Sharifi, M., Fathy, M., Mahmoudi, M.T.: A Classified and Comparative Study of Edge Detection algorithm. In: Proceeding of International Conference on Information Technology Coding and Computing (ITCC 2002). IEEE (2002)
5. Chidiac, H., Ziou, D.: Classification of Image Edges. In: Vision Interface 1999, Troise-Rivieres, Canada, pp. 17–24 (1999)
6. Ahmed, M.B., Choi, T.S.: Local Threshold and Boolean Function Based Edge Detection. IEEE Trans. Consumer Electronics 45(3) (August 1999)
7. Heath, M., Sarker, S., Sanocki, T., Bowyer, K.: Comparision of Edge Dectors: A Methodol-ogy and Initial Study. In: Proceeding of CVPR 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 143–148 (1996)
8. Canny, J.: A Computational Approach to Edge Detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI–8(6), 679–698 (1986)
9. Haraick, R.M., Shapiro, L.G.: Computer and Robot Vision, vol. 1. Addition- Wesley Pubishing Company Inc. (1992)
10. Marr, D., Hildreth, E.: Theory of Edge Detection. Proceedings of the Royal Society of London. Series B, Biological Sciences 207(1167), 187–217 (1980)

Segmentation of Mango Region from Mango Tree Image

D.S. Guru and H.G. Shivamurthy

Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore-570006, Karnataka, India
dsg@compsci.uni-mysore.ac.in, shiv_hg@rediffmail.com

Abstract. In this paper we propose a novel framework for segmentation of mango regions from its tree image. The proposed framework consists of mango localization followed by mapping of boundary information to the located region for segmentation. Initially thresholding is applied to each individual color band R,G and B by adaptive thresholding and later they are combined back. Application of smoothing and binarization to the combined image gives the location of mangoes along with noise. The texture features are extracted from each location then matched with template stored in the database to eliminate the noisy regions. Finally, locations of the mangoes are obtained and edge information is superimposed on to those locations for segmentation. An experiment is performed on our own dataset and efficiency is evaluated by computing the precision, recall and F-measure with respect to the human segmented images considering as a ground truth.

Keywords: Precision agriculture, Segmentation, Mango localization, thresholding, texture features.

1 Introduction

Modernizing agricultural practices with a help of an emerging technology leads to the environmental and economic sustainability with an optimized input in production of agricultural products. The process of identification and interpretation of in-field spatial variability through information and technology for effective management of agricultural practices such as soil mapping, disease mapping, weed mapping, selective harvesting and quality analysis, is known as precision agriculture (PA). Harvesting is the final stage of any agricultural practice. Manual process of selective harvesting will consume more time, manpower and not accurate. Selective harvesting is very popular today because we harvest only the matured crop, and while harvesting we also grade the crop based on the maturity level noticed. The clare valley and Margaret river regions of Australia is famous for wine grape cultivation. Bramley et al. [1] state that economic benefits that may accrue to grape growers and winemakers in Australia through the adoption of selective harvesting. Mango is one of the commonly cultivated commercial crops in many countries including India. Till today, the process of mango harvesting and grading is done manually and it consumes much of human effort and time. Nowadays, researchers are trying to automate the agricultural

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 201–211, 2013.
© Springer International Publishing Switzerland 2013

activities through the help of computer vision techniques, as it is very helpful to the formers, dealers and consumers. The mango fruits should be harvested at green mature stage. Selective harvesting of green matured mangos is helpful for maturity grading by distinguishing ripe and unripe mangos so that ripe mangos are used at the earliest and unripe mangos can be stored for some time. Further, quality wise grading of mangos can also be done during selective harvesting.

So, given an image of a mango tree, the problem of selective harvesting of mangos can be defined as a problem of localization and segmentation of mango regions from it. On the other hand it contributes the problem of spatial correlation since both leaf and fruits are green in color. Mango region segmentation is a process of acquiring knowledge about mango regions in an image. It is a very first task in a mango image analysis process such as disease mapping, variable spraying and selective harvesting. Quality of the subsequent tasks will depend on the success of the mango region segmentation process. Color and texture are the two essential features in mango tree image analysis. Texture is an efficient measure to estimate the structure, orientation, roughness, smoothness, and regularity differences of mango regions in the mango tree image. Two usual problems in mango region segmentation are over-segmentation: where an image is segmented into a more number of regions than the actual mango regions in the mango tree image and under-segmentation: where an image is segmented into a less number of regions than the actual mango regions in the mango tree image. The over-segmentation and the under-segmentation of mango regions happen usually in case of images with spatially varying illumination. Also, the leaves are occluded on mangoes in the mango tree image. Hence balancing the over and under segmentation in unconstraint environment of the mango tree image is a challenging issue in the area of color image segmentation.

The automation of agricultural practices for significant increase in food production is emerging out as a new challenge for computer vision community. Ducournau et al. [2] proposed a machine vision approach to count the number of emergent radical tips on seed-lots, under the constrained environment and segmentation is accomplished with the help of thresholding and morphological operators. Green vegetation region segmentation [3] using the IHS (Intensity, Hue, Saturation) and RGB (Red, Green, Blue) color space for color feature extraction and then apply mean shift and BPNN (back propagation neural network) for segmentation. One of the central point of precision agriculture is the selective treatment of weeds [4] and it is achieved in three different stages at which each different agricultural elements is extracted. They are segmentation of vegetation against non-vegetation, crop row elimination and weed region extraction. The segmentation is done by thresholding on the basis of dominant G-component, dominated B-component, minimum and maximum intensity in the image. Segmentation of lesquerella flowers was proposed by Thorp et al. [5] based on thresholding the images in HIS color space and boundary conditions using six parameters, including the maximum and minimum hue, saturation and intensity in image. The popular general unsupervised segmentation of color-texture region in image [6] is presented and it is named as J-seg which consists of two independent steps: color quantization and spatial segmentation.

All the segmentation problems which are related to a precision agriculture are considered to be two class problems (the pixel belongs to vegetation/non-vegetation).

In this work, we made an attempt to classify pixel belongs to a region of mango or non-mango in a mango tree image. In our framework we use texture features for template matching and color features for sensing the object of our interest. The proposed framework consists of two major phases they are:

(1) Mango localization, and
(2) Mapping the edge information for segmentation.

The rest of the paper is organized as follows. In section 2, we present the overview of the proposed method. The experimentation and results are described in the section 3. Finally paper ends with conclusions in section 4.

2 Proposed Framework

In this section, we propose a new framework for mango region segmentation, which consists of two phases namely, mango localization and edge mapping. In first phase, we apply thresholding to each color band separately then combined them. The combined image (T-image) is said to have j-number of regions. These regions are considered as mango regions. We perform smoothing to eliminate the non-mango regions in the thresholded image. Then apply binarization to the resultant image to get B-image. The B-image has say k-number of regions assumed to be the mango regions. Finally we achieve mango localization through template matching which eliminates the non-mango regions in the B-image and results with an image (L-image) having m-number of mango regions where $j \geq k \geq m$. In second phase, the edge information from the original image is superimposed on to the regions of the L-image for segmentation. The block diagram of mango region segmentation is shown Fig .1.

2.1 Image Thresholding

The reflectance on the surface of mango regions is higher than that of non-mango regions, so higher intensity will be preserved by the regions belonging to mangos. This is an important clue for us to distinguish mango regions from non mango regions. Based on this clue we apply a simple thresholding to segment the mango regions from its tree image. We separately threshold each R, G & B components in the mango tree image and then combine the resultant images back (T-image). The middle value of the intensity range in a given image intensity distribution will be taken as threshold. The thresholding process can be formulated as shown in Eq. (1) and (2).

$$T(i,j)=\begin{cases} A(i,j), A(i,j) \geq \text{threshold} \\ 0, \quad A(i,j) < \text{threshold} \end{cases} \tag{1}$$

$$\text{Where} \quad \text{Threshold} = \frac{\text{Max}(\text{intensity}) - \text{Min}(\text{intensity})}{2} \tag{2}$$

Here A is the mango tree color image and T is the thresholded image. Fig.2a-2d depict the original color image and R, G, B, components of the color image given.

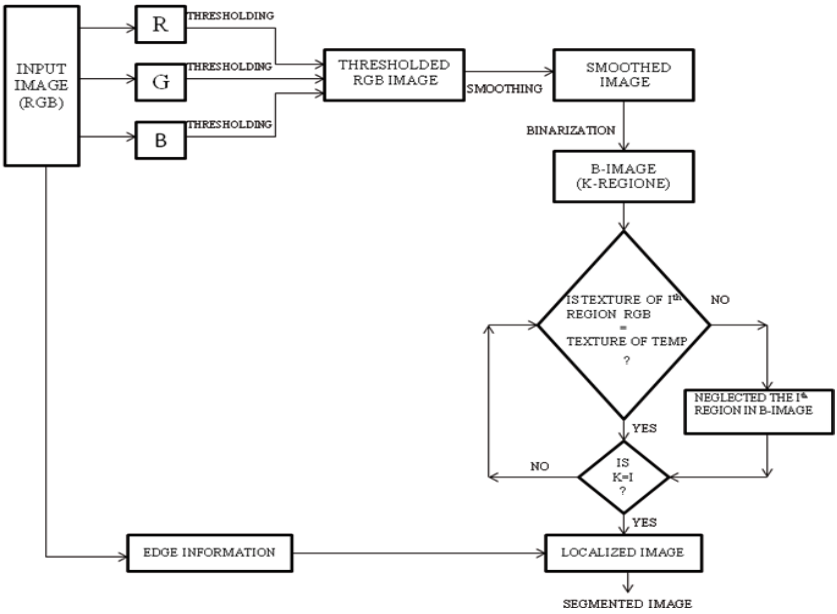


Fig. 1. Block diagram of the proposed framework

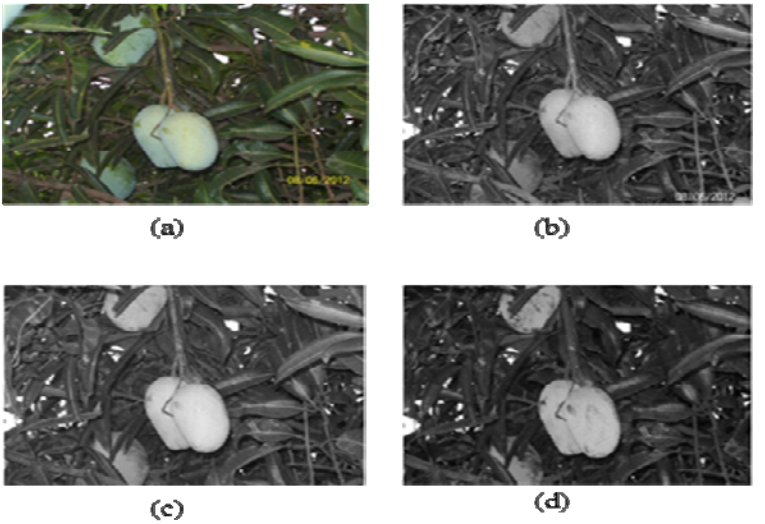


Fig. 2. (a) A given mango tree image (RGB). (b) The R-component of the input image. (c) The G-component of the input image (d) The B-component of the input image.

The major difficulty in mango tree image segmentation is varying illumination which can be seen in the mango regions at upper left corner, upper middle and lower middle of the Fig.2a. Initially we assume the regions in the thresholded image as mango regions. Fig.3a-3d depict the thresholded images of each R, G, B component and combined image of the original color image respectively.

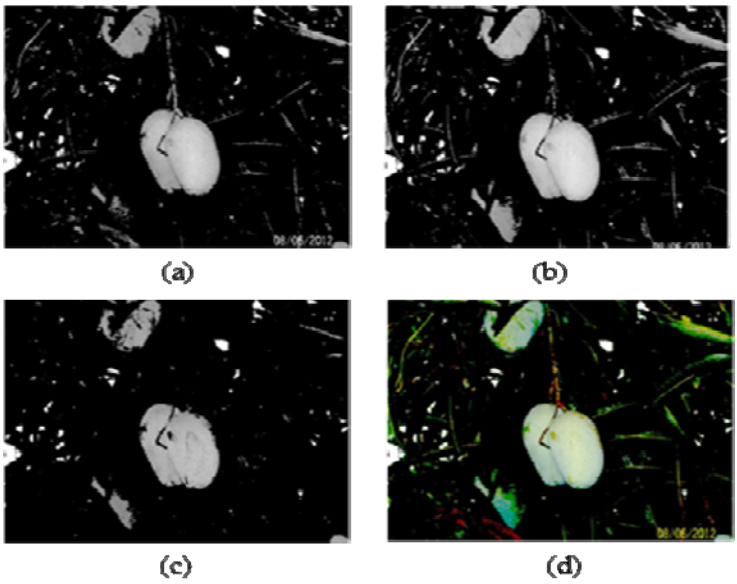


Fig. 3. (a) Thresholded image of R-component. (b) Thresholded image of G-component. (c) Thresholded image of B-component. (d) Thresholded image of given input image Fig.2a.

2.2 Smoothing and Binarization

The small regions and sharp transitions in the thresholded image are non-mango regions, because the mango regions are relatively larger regions. The sharp transitions in intensity levels can be seen only when there is random noise in the image. So, we perform smoothing to remove small regions and sharp transitions (non-mango regions) in the thresholded image of each of the R,G and B components separately and combined them. We use a 3x3 mask for smoothing because it will not affect lager mango regions and also it results in less blurring effect. The mango region segmentation is a two class problem, we convert each smoothed R, G and B component into binary images where, white pixels represent mango regions and black will represent non-mango regions. Then combine the resultant images into a single binary image (B-image) as shown in the Fig.4. The binarization process will fixes up a contrast break-point between pixels belonging to mango regions and that of non-mango regions.

Fig. 4. Binarized image (B-image)

2.3 Template Matching

Here we use template matching technique to eliminate the non-mango regions in the B-image. For this purpose, we extract the texture properties of the original image from the regions located in the B-image and match with the templates stored in the database. We retain only those regions whose texture properties are similar to mangoes template. The template of mango is as shown in the Fig.5.

Fig. 5. Few templates of mangoes

Texture features are the very fundamental and invariant properties of images. Each image has its own texture properties which describe different image regions present in it. In the image classification and image segmentation literatures texture properties of images have been efficiently used. Statistical texture features proposed by Haralick et al. [7] are used for template matching to eliminate the non-mango region in the binarized image (B-image). Initially Gray Level Co-occurrence matrix (GLCM) is computed for the gray image using the pair-wise occurrences of image resolutions. And the various texture properties are calculated using the GLCM obtained. In our frame work three different texture features are used. Let us assume that P is the gray level co-occurrence matrix obtained from the image region rx, expressions for different texture features which we have used are as follows.

(1). Contrast:

$$f_1 = \sum_{n=0}^{Ng-1} n^2 \left\{ \sum_{i=1|i-j|=n}^{Ng} \sum_{j=1}^{Ng} P(i,j) \right\}$$

(2). Correlation:

$$f_2 = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

(3). Entropy:

$$f_3 = -\sum_i \sum_j p(i,j) \log(P(i,j)).$$

Notation:

$p(i, j)$ is the $(i, j)^{th}$ entry in a normalized gray-tone spatial dependence matrix.

μ_x , μ_y , σ_x and σ_y are the means and standard deviation of p_x and p_y.

N_g is the Number of distinct gray levels in the quantized image.

Below is the algorithm for texture based template matching.
Algorithm: template matching.

Input: B-image (k-regions).
Output: mango localized image
Method:

1. For I^{th} region in the B-image, extract the corresponding region in the original given image.
2. Compute the texture features (Contrast, Correlation and Entropy) of the region extracted.
3. Compute the dissimilarity between texture features extracted from obtained region and templates stored in the database.
4. If dissimilarity is less than the threshold, then region belongs to mango otherwise it is non-mango region, neglect it.
5. Repeat the above procedure till (I=k).

End of algorithm.

The output image of the above algorithm gives location of mangoes. To accurately show the boundaries of each mango located and also to clearly distinguish the mangos in the cases of occlusion, we extract the edges from located regions of the original image and superimpose on to the localized image for segmentation. We use canny edge detection operator for edge extraction. Finally we obtain segmented image of the original image Fig.2a is shown in the Fig.6.

Fig. 6. Segmented image

3 Experimentation and Results

3.1 Dataset

Since, we could not find any dataset with mango tree image in the literature, we have captured 44 natural mango tree images in a natural lighting condition using Kodak digital camera (8.2-megopixel) with resolution 480x480dpi from different directions and with the distance of less than five feet because, the image intensity at pixel depends on the optical properties of the surface material, the surface shape and spatial distribution of the incident illumination. The images we captured consists mango leaves and branches in addition to mangos. We captured images under different illumination conditions and occlusion of leaves on the mangoes. Sample images of our own dataset is show in the Fig.7

(a) (b) (c) (d)

Fig. 7. Sample images of mango trees

3.2 Results

The quantitative results of mango region segmentation are computed on our own data set using precision, recall and F-measures as given below,

$$\text{Precision(P)} = \frac{\text{MRS}}{\text{MRS+NMRS}} \tag{3}$$

$$\text{Recall(R)} = \frac{\text{MRS}}{\text{AMR}} \tag{4}$$

$$\text{F–measure} = \frac{2*\text{P}*\text{R}}{\text{P+R}} \tag{5}$$

Here MRS is the mango region segmented correctly, NMRS is the non mango region segmented as a mango region and AMR is actual mango regions present in the image (ground truth) and human segmented image considered as a ground truth. The more segmented results are shown in Fig.8.

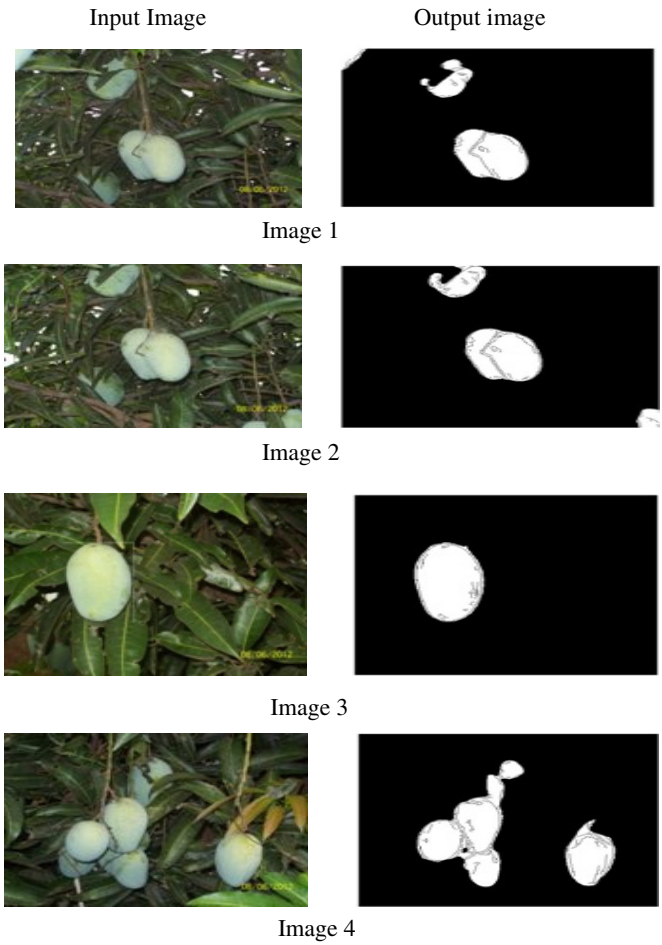


Fig. 8. Segmented mango tree images with original input images

In the image (3) we exactly segment the mango region, but in the case of image (1), image (2), and image (4) there is a over segmentation due to the occlusions of leaf on the mango which is still a challenging issue in mango region segmentation. The graphical representation of precision, recall and f-measures obtained for our data set is as shown below Fig.9, Fig.10 and Fig.11 respectively.

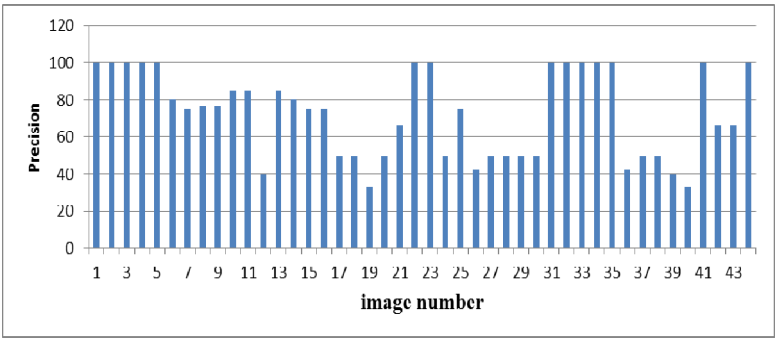


Fig. 9. Precision of the proposed method

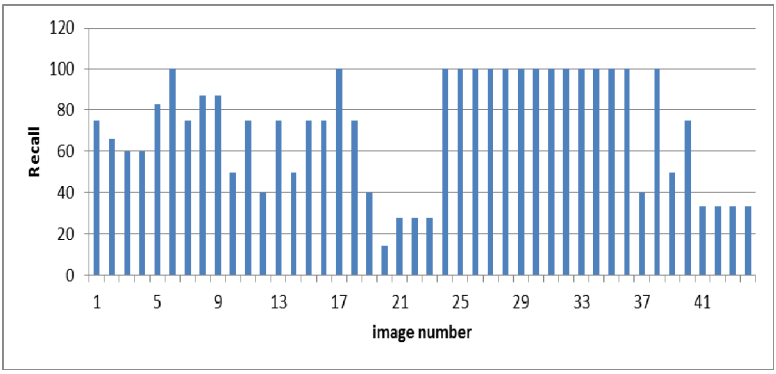


Fig. 10. Recall of the proposed method

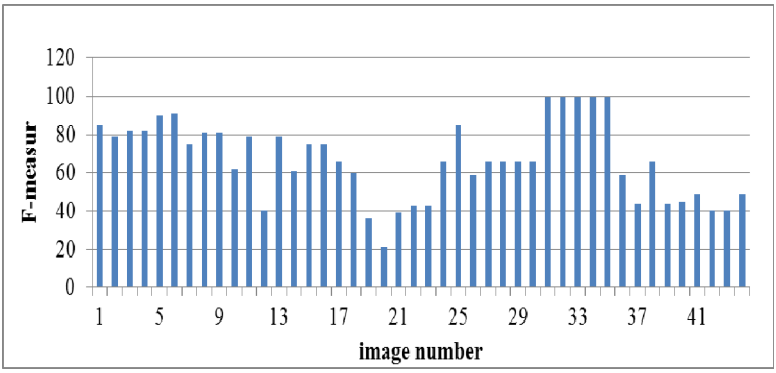


Fig. 11. F-measure of the proposed method

4 Conclusion and Future Work

In this paper, we have proposed a novel approach for mango region segmentation from the mango tree image. The intensity of the pixel location depends on the optical property and shape roughness of the objects present. Texture is the only feature to achieve discrimination between objects in an image. Through this framework we have exploited the intensity distribution and texture features of the objects in images for the purpose of segmentation. The segmentation of mango regions from its tree image taken in an unconstrained environment is a very challenging task. Presently, there are no mango tree datasets available in the literature and hence we have created our own dataset consisting of 44 mango tree images taken in an unconstrained environment. The quantitative performance of the proposed method is analyzed and tabulated using Precision, Recall and F-measures. We have achieved 72.77 % of Precision, 71.43 % of Recall and 66.70% of F-measure. The poor results (image-20) are evidence in the Fig.11 due to the problem of illumination, leaves shadows occluded on the mango regions. In future, we would like to enhance the efficiency and accuracy of the proposed method by making better use of texture features and with the addition of features such as shape, spatial geometry and solidity etc.

References

1. Bramley, R.G.V., Proffit, A.P.B., Hinze, C.J., Pearse, B., Hamilton, R.P.: Generating benefits from precision viticulture through selective harvesting. In: Proceedings of the 5Th European Conference on Precision Agriculture, pp. 891–898 (2005)
2. Ducournau, S., Feutry, A., Plainchault, P., Revollon, P., Vigouroux, B., Wagner, M.H.: An image acquisition system for automated monitoring of the germination rate of sunflower seeds. Computers and Electronics in Agriculture 44, 189–202 (2004)
3. Zheng., L., Zhang., J., Wang, Q.: Mean-shift-based color segmentation of images containing green. Vegetation Computers and Electronics in Agriculture 65, 93–98 (2009)
4. Burgos-Artizzu., X.P., Ribeiro., A., Tellaeche., A., Pajares., G., Fernández-Quintanilla, C.: Analysis of natural images processing for the extraction of agricultural elements. Image and Vision Computing 28, 138–149 (2010)
5. Thorp, K.R., Dierig, D.A.: Color image segmentation approach to monitor flowering in lesquerella. Industrial Crops and Products 34, 1150–1159 (2011)
6. Deng, Y., Manjunath, B.S.: Unsupervised Segmentation of Color-Texture Regions in Images and Video. Pattern Analysis and Machine Intelligence 23, 800–810 (2001)
7. Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture Feature For Image Classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3, 610–621 (1973)

Detection and Removal of Scratches in Images

S. Bhuvaneswari, T.S. Subashini, and N. Thillaigovindan

bhuphdofficial@gmail.com,
rtramsuba@gmail.com,
thillai_n@sify.com

Abstract. In this paper a detection/restoration method to detect and remove line scratches in still images, regardless of their orientation, colour, and shape is presented. In this work the two properties of scratches are considered namely: scratches have high contrast compared with its neighbours and they usually occur vertically and it is more than half of the image in length are taken into account. Firstly a simple thresholding technique is applied for detecting the candidate scratch pixels. The pixels detected in the first step are then used as a mask for removing the scratch during the restoration step. The detected scratch is inpainted using our higher order non-adaptive interpolation approach based on a 13x13 neighbourhood. The experimental result shows that the proposed method works well for both simple and complex scratches that are present in uniform or less complex backgrounds. When experimented with complex backgrounds the interpolation artefact namely blurring becomes pronounced. The proposed work can be effectively used to automatically detect and remove the scratch from uniform and less complex static images without user intervention in any stage of the process.

Keywords: Scratches, detection, inpainting mask, inpainting algorithm, image restoration.

1 Introduction

The image can be understood as a two dimensional function f(x, y) where x and y are spatial coordinates, and the amplitude of f at any pair of coordinates(x, y) is called the intensity or grey level of the image at that point [1].Scratches are defects in old films, and images that cause damage to the clarity of the images. Scratches are a kind of anisotropic interference which occurs vertically in common. Scratches in images are visible as vertical lines which are having bright or dark intensity, when compared with the neighboring pixels. Therefore the scratches are categorized by considering their length, and luminance [4]. According to the length of the scratch, it can be classified into two types namely,

a) Principal scratch b) Secondary scratch

Principal scratch: The length of this type of scratches is more than 95% of the image height.

Secondary scratch: This type of scratches is very short compared to principal scratches.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 212–223, 2013.
© Springer International Publishing Switzerland 2013

According to the luminance, the scratches are categorized as follows.

Positive scratch: The scratch occurs in dark intensity neighbourhood.

Negative scratch: The scratch occurs in light intensity neighbourhood.

Fig. 1 shows the example of degraded regions due to principal and secondary type of scratches.

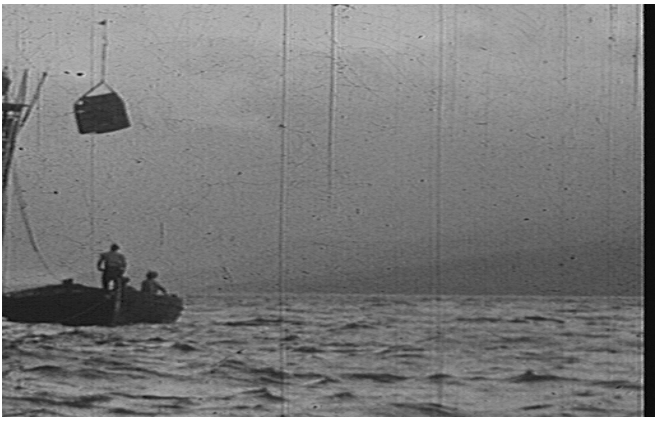


Fig. 1. Image with both types of scratches

Taking the characteristics of scratches [5] into consideration, this paper, presents a new scratch detection method, based on thresholding and scratch selection criteria.

Nowadays, digital image inpainting has been widely researched in the field of digital image processing. The term inpainting [2] refers to automatic filling of a specific region in an image called as a target region (mask). Usually the target is selected manually and marked as a mask, which will be inpainted based on the details obtained from the surrounding pixels.

It can be seen from Fig. 2 how the tennis racket in the original image has been removed after manually selecting the tennis racket as the target region.

There are many areas where inpainting is used. Applications of inpainting include restoration of photographs, films, removal of texts, logos, stamps, scratches, red eye etc.

The common requirement for most of the image inpainting algorithms in [2][19][20][21] is that the region to be inpainted is manually selected by the user. As a first step the user manually selects the portions of the image that require restoration. This is usually done as a separate step and involves the use of other image processing tools. Then image restoration is done automatically, by filling these regions with the information obtained from the surrounding pixels (or) from the whole image.

This paper aims in detecting and removing scratches automatically without any manual intervention. This approach compares the intensity values of the neighbourhood pixels, and based on the threshold, the scratched pixels are identified. After filtering, noise free scratches are selected in the first step. In the second step the detected scratch is considered as a target region for inpainting. Original image is compared with a mask and the neighbouring pixels of the scratch are extracted. A 13x13 pixel neighbourhood is used for extraction. The extracted values are summed and the result is averaged. Then the new value is replaced in the region of scratch. The process is repeated until the scratch inpainted fully.

a) Original image

b) Manually selected (tennis racket) mask to be inpainted

c) Inpainted tennis racket

Fig. 2. Inpainting process

The rest of the paper is organized as follows. Section 2 reviews the related work in this area. Section 3 presents the proposed methodology. Section 4 gives the experimental results and section 5 concludes the paper.

2 Related Work

This section gives an overview of the work that has been carried out in the area of automatic scratch detection and inpainting of scratches. Scratches differ by size, orientation, colour etc. and automatic detection of scratches is very challenging.

Kokaram's spatial model for scratch line detection is generalised by Vittoria Bruni and Domenico Vitulano [6]. Nie Shilang et al. [7] have proposed a vertical scratch detection algorithm based on edge detection. Kyung-tai Kim et al. [8] have proposed a neural network based detection method and genetic optimisation algorithm for restoration. A simple 1D- extrema detector for identifying line scratches and a kalman filter to reject the false scratches is proposed by Laurent Joyeux et al. [9].They have also proposed a Bayesian restoration technique for damaged areas. Francesco Isgro and Domenico Tegolo [10] have approached the problem of scratch restoration as a optimization problem using a distributed genetic algorithm. Nam-Deuk Kim and Satish Udpa [11] have proposed a non-linear edge detection and line scratch removal method considering the differences and sums of four neighbour pixels. The method proposes the use of maximum or minimum operators in the neighbourhood. Zeng Qingyue and DingYoudong [12] have put forward a scratch line detection and restoration algorithm based on canny operator. Ardizzone et al. [13] have given a method for the analysis, detection, and restoration of line scratches in still images, irrespective of their orientation, colour, and shape. Zhang Hongying et al. [14] have proposed a method to enhance the scratches and then detect them using the information of the frames. They have also approached the scratch removal as an inpainting problem using p-Laplace operator. Inpainting and the fundamental problem of image processing where studied by Jianhong Shen [15] from the point of view of Partial Differential Equations. An analytical study of different image inpainting techniques is carried out by Supriya Chhabra et al. [16]. An image inpainting technique based on 8-neighborhood fast sweeping method is developed by Sagar and Kashif Hussain [17]. A computationally efficient algorithm using successive elimination in 8-pixel neighbourhood for inpainting is developed by Muthukumar et al.[18]. A fast exemplar based approach for filling the missing portions in an image is developed by Bhuvaneswari et al. [19]. This technique can also be used for restoring old photographs, damaged films and for removing superimposed text like date, sub titles etc. A hybrid region filling algorithm composed of a texture synthesis technique and an efficient interpolation method with a refinement approach is developed by Han-Jen Hsu et al. [20]. An algorithm for digital inpainting using fast marching method for level set applications is developed by Alexandru Telea [21].

In this paper an automatic scratch detection and inpainting technique with the focus on improvement in automatic detection and removal of scratches without user intervention.

3 Methodology

The proposed work is done in two steps.

- Detection of scratches
- Inpainting

First step detects the area where the scratch is located without user interruption. In the second step the scratch detected in the first step is used as a mask for inpainting. In this step the image region is inpainted using the proposed inpainting method which is based on neighbourhood pixel grey values. The flow diagram of the proposed work is shown in Fig. 3.

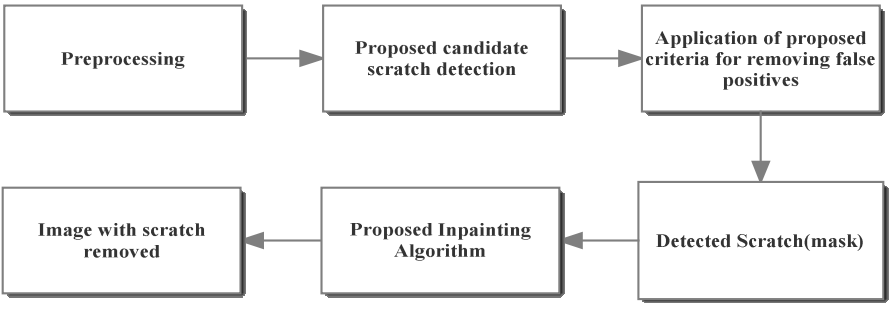


Fig. 3. Flow diagram of the proposed work

3.1 Preprocessing

In the preprocessing step the image is converted into gray scale image and the noise is removed by applying M3 filter.

M3 filter is a smoothing filter. It is a hybrid of mean and median filter and it is widely used to remove noise and, the subjective quality of this M3filtered image is far better than median and mean filter [22]. So in this work M3 filter is proposed. It replaces the central pixel by the maximum value of mean and median for each sub images. The Fig. 4 shows the preprocessing results.

3.2 Detection of Candidate Scratches

Normally the width of the scratch ranges from 3 to 10 pixel width [3]. As the intensity of the scratch pixels is higher than its neighborhood pixels, the pixels having intensity values beyond certain threshold is found detected scanning from left to right, the adjacent pixels whose width that lies in the range of 3 to 10 are selected as candidate scratch pixels. At the completion of the process the resultant binary image will have

Fig. 4. Pre-processing steps

a) RGB colour image b) Gray scale image c) M3 filtered image

scratch pixels with some noise. A scratch pixel is normally an edge pixel and the intensity of edge pixels is usually high. The threshold has been set as 160 in this work as all scratch pixels are usually above this intensity value.

3.3 Scratch Detection without Noise

Before creating a mask, the small objects which are not satisfying the criteria to be a scratch are removed from the binary image. The objects that have length less than 50 pixels, the value which is found through experimentation, are removed. Now, the resultant binary image which contains the scratches is taken as the mask for inpainting.

3.4 Inpainting Using Neighborhood

This step involves the computation of the grey value of a particular pixel in the image using the grey values of its neighbouring pixels.

- It is a spatial domain technique which is used for image enhancement in common.
- The pixel value f (x,y) under consideration is computed based on the pixel values of either a 3x3, 5x5 or 7x7 neighbourhood depending on the need.
- The step by step approach for inpainting is given in Fig 5.
- In this paper a 13x13 pixel neighbourhood is taken for computation, as the width of the scratch ranges from 3 to 10 pixels and it is given by

$$f(x,y) = \frac{1}{168}\sum_{i=-6}^{6}\sum_{j=-6}^{6}\begin{bmatrix} f(x-i,y+j)+f(x+i,y+j)+ \\ f(x-i,y-j)+f(x+i,y-j) \end{bmatrix}$$

$$with \cdots f(x,y) = 0$$

- The average of the 13x13 neighbourhood is found out leaving out the centre (scratch pixel). So the denominator in the expression is 168 and not 169. The scratch pixel is replaced with average value computed.
- The process is continued until all the scratches are inpainted.

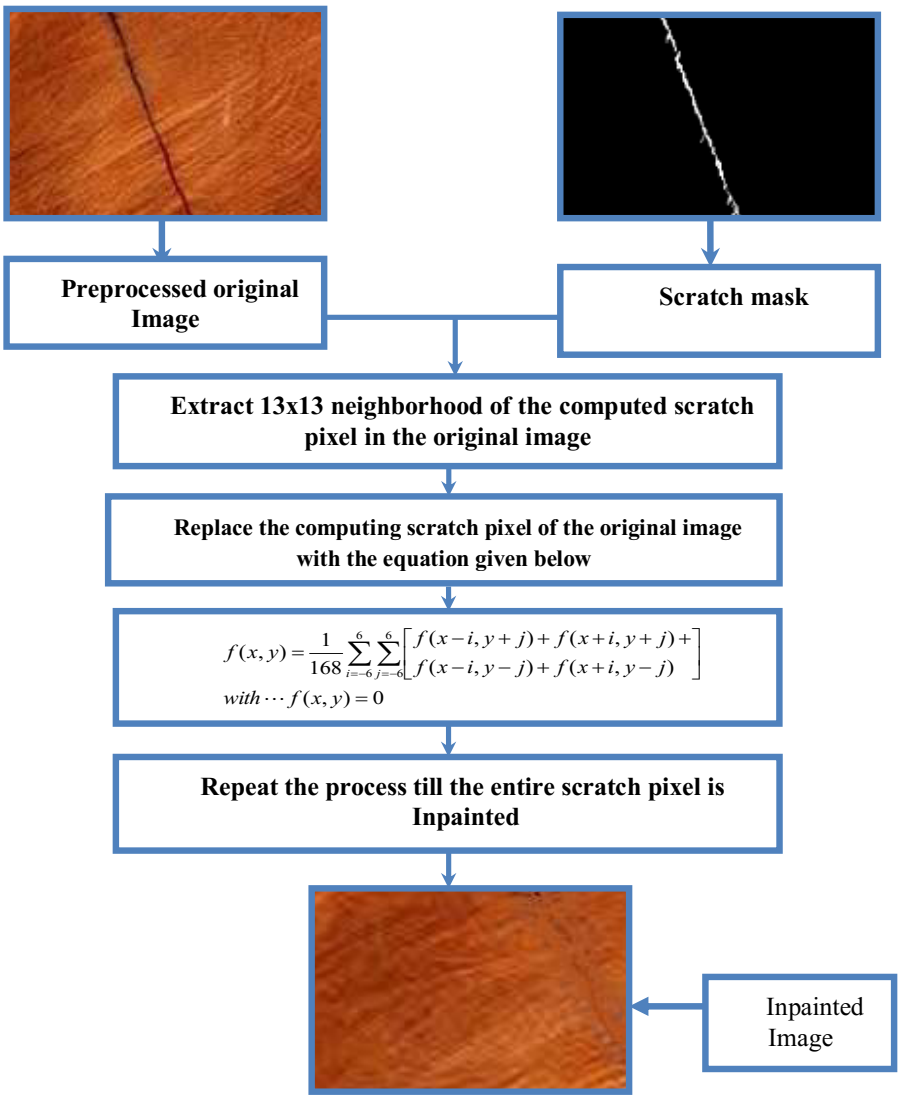


Fig. 5. Block diagram of the inpainting step

4 Experimental Results

In order to prove the performance of our proposed scratch detection algorithm, in this section, the large number of still images is experimented. The images are taken from

http://www.google.com/linescratches, http://www.mee.tcd.ie/~ack/cd/lines/lines.htm. The following are the images are considered to evaluate our algorithm. The images with simple and complex background are given as input to the proposed system and the outcomes are shown in Figures 6, 7, 8.

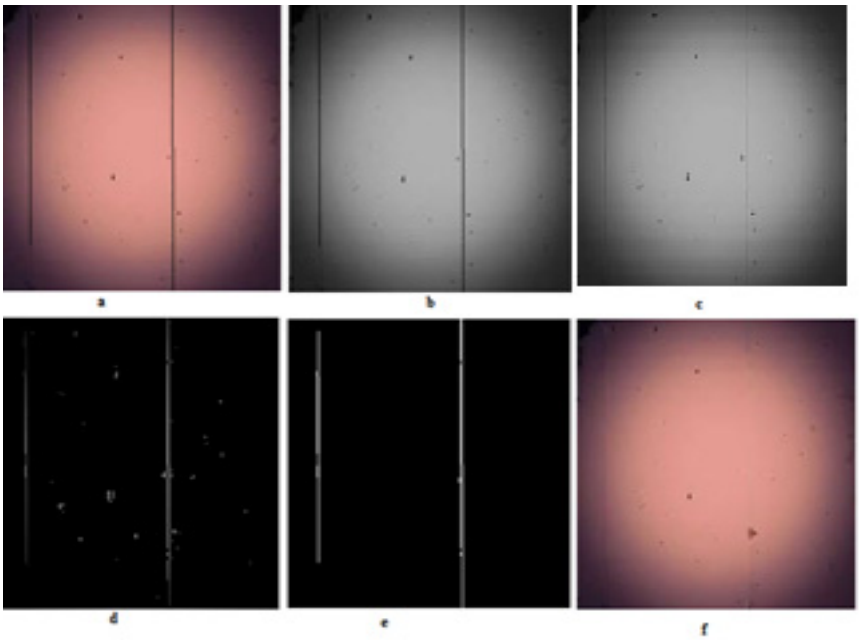


Fig. 6. Detection and removal of scratch in the image a) RGB color image b) gray scale image c) M3filtered image d) scratch detected with noise e) scratch noise eliminated f) image without scratch

The image in Fig. 6 is used to test the proposed method. The image consists of noise and two scratches. Fig. 6(d) shows the detected scratch and Fig. 6 (e) shows the removed scratch inpainted leaving the texture without damage.

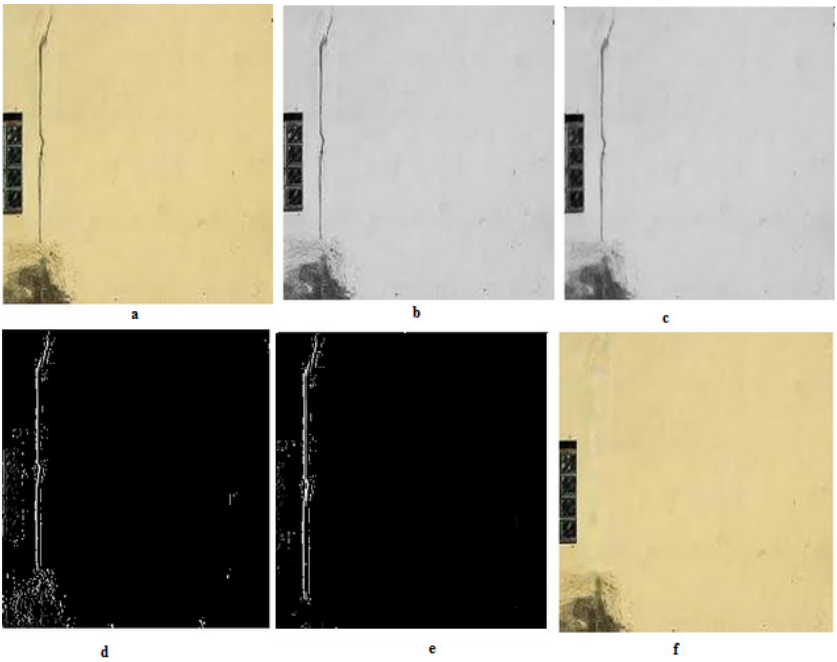


Fig. 7. The crack in the wall is detected and inpainted a) RGB color image b) gray scale image c) M3 filtered image d) scratch detected with noise e) scratch detected with noise eliminated f) image without scratch

In the Fig. 7 the crack in the wall is detected and inpainted automatically without user intervention. The proposed method is restricted to detect scratches only, so that a blotch which is present at the bottom left corner is not considered as a scratch. Fig. 7(f) shows the scratch inpainted with the background texture information.

In Fig. 8 the scratch is seen at the middle of the tile. The proposed algorithm was used to detect the crack and inpaint it automatically. Fig. 8 (f) shows the inpainted tile.

The graph in Fig. 9 shows the PSNR values calculated using the proposed method and exemplar method for the images shown in Fig. 6, 7, 8. The quality improvement is achieved by the proposed method is evident from the graph.

The proposed method detects the scratch automatically and the detected region is used as a mask for inpainting. This method uses 13x13 neighborhoods to inpaint the mask region.

However in exemplar based method the mask (i.e.) the scratch to be inpainted has to be selected manually. It uses the most probable neighborhood. The PSNR values shown in Fig. 9 indicate that the proposed method can be effectively used in detecting and inpainting scratches in images.

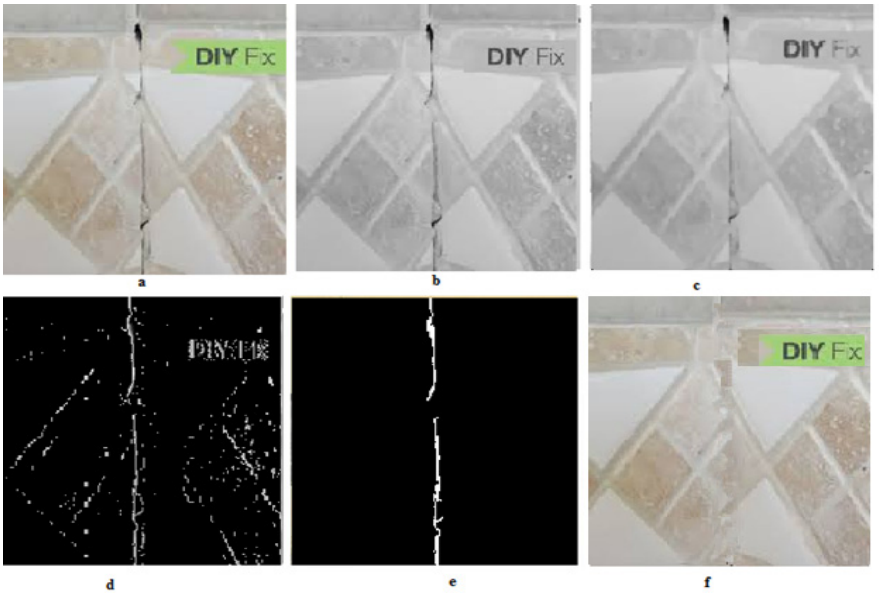


Fig. 8. The scratch on the tiles detected and inpainted

a) RGB color image b) gray scale image c) M3 filtered image d) scratch detected with noise e) scratch with noise eliminated f) Image without scratch

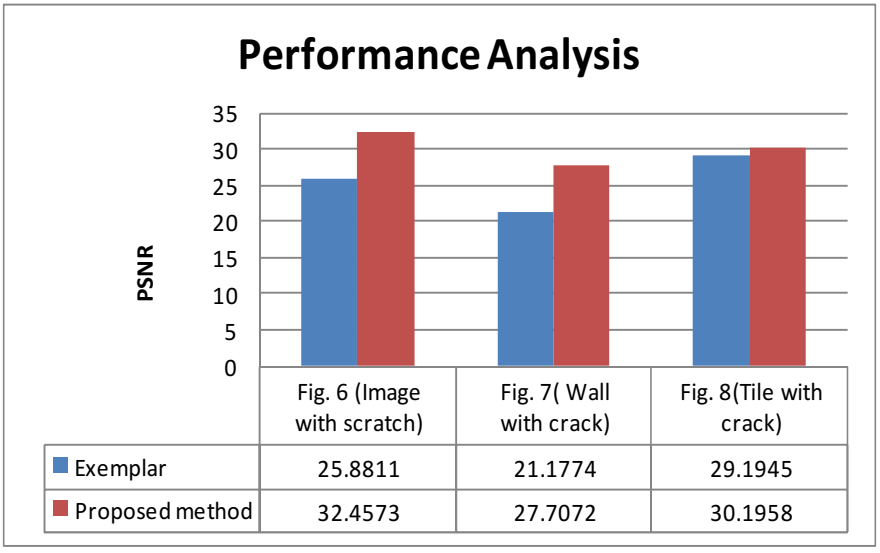


Fig. 9. Comparative analysis of proposed method with criminisi's exemplar method

5 Conclusion

In this paper, a new method is introduced to detect scratches and remove it automatically both in colour and grey images. The detected scratch in image is taken as a mask for inpainting. Scratches have high contrast compared with its neighbours. Also scratches usually occur vertically and it is more than half of the image in length. These two properties are taken into an account. The detected scratch is inpainted using our new method which uses 13x13 neighbourhood method. Experimental results have demonstrated that the proposed method can be effectively used to detect the scratch without user interruption. During the inpainting stage the detected mask is taken as a target region for inpainting. The proposed method works well for both simple and complex scratches with uniform or less complex images and it can be used to create automatic detection of scratch masks and to remove the scratch from the images without user intervention in any stage of the process. As future work, efforts are underway to apply this new method to detect scratches of shorter extent, as well as curvilinear ones.

References

1. Gonzalaz, R.C., Redwoods: Digital Image Processing, 2nd edn. Pearson Education (2002)
2. Criminisi, A., Perez, P., Toyama, K.: Region Filling and objects removal by exemplar based image inpainting. IEEE Transactions on Image Processing 13(9), 1–7 (2004)
3. Muller, S., Buhler, J., Weitbruch, S., Thebault, C., Doser, I., Neisse, O.: Scratch Detection Supported By Coherency Analysis of motion Vector Fields. In: ICIP, pp. 89–92 (2009)
4. Kim, K.-T., Kim, E.Y.: Film Line scratch Detection using Neural Network and Morphological Filter. In: CIS, pp. 1007–1011 (2008)
5. Malvia, A.: Scratch Detection and Removal in Motion Picture Images. In: IET International Conference on Visual Information Engineering, pp. 99–104 (2006)
6. Bruni, V., Vitulano, D.: A Generalized Model for Scratch Detection. IEEE Transactions on Image Processing 13(1) (2009)
7. Shiliang, N., Hongying, Z., Liping, Z., Yang, F., Brost, V.: Vertical Scratches Detection based on Edge Detection for Old Film. In: IIS, pp. 257–259 (2010)
8. Kim, K.-T., Kim, B., Kim, E.Y.: Automatic restoration of scratch in old archieve. In: IEEE proceedings on the International Conference on Pattern Recognition, pp. 468–471 (2010)
9. Joyeux, L., Boukir, S., Besserer, B.: Film Line Scratch Removal using Kalmar Filtering and Bayesian Restoration. In: IEEE Workshop on the Application of Computer Vision, pp. 1–6 (2000)
10. Isgro, F., Tegolo, D.: Restoration of vertical line scratches with distributed genetic algorithm. In: IEEE Workshop on Computer Architecture for Machine Perception, pp. 249–254 (2005)
11. Kim, N.-D., Udapa, S.: Nonlinear Operators for Edge Detection and Line Scratch Removal. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 4401–4404 (1998)
12. Qingyue, Z., Youdong, D.: Scratch Line Detection and Restoration Based on Canny Operator. In: IEEE Proceedings on the Asia-Pacific Conference on Information Processing, pp. 148–151 (2009)

13. Ardizzone, E., Dindo, H., Mazzola, G.: Multidirectional Scratch Detection and Restoration in Digitized Old Images. Eurasip Journal on Image and Video Processing 2010:680429, 1–11 (2010)
14. Yadong, W., Zhonglin, K.: An Efficient Scratches Detection and Inpainting Algorithm for Old Film Restoration. In: IEEE Proceedings on International Conference on Information Technology and Computer Science, vol. 1, pp. 75–78 (2009)
15. J., Shen (Jackie), J.: Inpainting and the Fundamental Problem of Image Processing. SIAM News 36(5) (2003)
16. Chhabra, S., Lalit, R., Saxena, S.K.: An analytical Study of Different Image Inpainting Techniques. Indian Journal of Computer Science and Engineering (IJCSE) 3(3), 487–491 (2012)
17. Sagar, G.V.R., Kashif Hussain, S.: An image inpainting Technique based on 8-neighborhood Fast Sweeping Method. International Journal of Computer Science and Technology (IJCST) 2(3), 100–103 (2011)
18. Muthukumar, S., Krishnan, N., Pasupathi, P., Deepa, S.: Analysis of Image Inpainting Techniques with Exemplar, Poisson. Successive Elimination and 8 pixel neighborhood methods 9(11), 15–18 (2010)
19. Bhuvaneswari, S., Subashini, T.S., Soundharya, S., Ramalingam, V.: A novel and fast exemplar based approach for filling portions in an image. In: IEEE Proceedings on the International Conference on Recent Trends in Information Technology (ICRTIT), pp. 91–96 (2012)
20. Hsu, H.-J., Wang, J.-F., Liao, S.-C.: A Hybrid Algorithm With Artifact Detection Mechanism for Region Filling After Object Removal From a Digital Photograph. IEEE Transactions on Image Processing 16(6), 1611–1622 (2007)
21. Telea, A.: An image Inpainting Technique Based on the Fast Marching Algorithm. Journal of Graphics Tools 9(1), 25–38 (2004)
22. Thangavel, K., Manavalan, R., Laurence Aroquiaraj, I.: Removal of Speckle Noise from Ultrasound Medical Image based on Special Filters: Comparative Study. ICGST International Journal on Graphics, Vision and Image Processing (GVIP) 9(3), 25–32 (2009)

An Automatic Method to Locate Tumor from MRI Brain Images Using Wavelet Packet Based Feature Set

T. Kalaiselvi and Karthigai Selvi

Image Processing Lab, Department of Computer Science and Applications
Gandhigram Rural Institute - Deemed University, Gandhigram, Tamilnadu, India
{kalaivpd,karthigachandru}@gmail.com

Abstract. This paper developed a fully automatic method to locate the brain tumor from Magnetic resonance imaging (MRI) head scans using wavelet packet transformation (WPT) based feature set. WPT is used to extract high frequency data from all sub bands of MRI images. Modulus maximum is used to detect singularities among these high frequency features and thus isolates the hyper intense nature of tumors. These tumor areas are detected by preparing a mask of modulated images and then compared it with the original scans. This method does not require any preprocessing operations like seed selection, initialization and skull stripped scans of existing methods. Experiments were done with the sample images collected from popular hospitals and clinics. The results were visually inspected for the outputs. The quantitative validation was done with the Chi-square test. It performed significance study to identify the goodness of fit, the probability of fitness is above 0.75.

Keywords: wavelet packet transformation, MRI, modulus maximum, segmentation.

1 Introduction

MRI images have high contrast and high resolution. Nowadays this scanning is suggested mostly to the brain disorder patients. It is a dynamic and flexible technology and more sensitive in detecting brain abnormalities during the early stages of disease. The two basic types of MRI images are T1 weighted and T2 weighted. The visual difference of those is cerebral spinal fluid (CSF). It is bright in T2 images and dark in T1. T2 weighted images are very sensitive in detecting brain pathology [1]. The skull and tumor cells are having same intensity so the existing algorithms need to remove the skull [2]. But in proposed work no need to remove the skull.

Uncontrollable growth of cells leads to tumor in brain, the pixel intensity of tumor and CSF are more or less same. So manual segmentation needed in MRI images to ensure the tumors. But MRI scanner produces hundreds of images per patient, the manual segmentation consumed more time. Hence automation in

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 224–233, 2013.
© Springer International Publishing Switzerland 2013

segmentation takes place. These segmentation methods are basically categorized as Region based and Boundary based methods. In region based method, k-means and fuzzy c means [2] are outperformed methods, but centre point initialization is critical. To overcome this drawback, knowledge based systems were added [3]. Some other algorithms proposed like Level-set evolution [4], multi spectral histogram [5] for region based segmentation. In Boundary based segmentation, snake model, curve fitting, fluid vector flow [6], active contour model [7] and level sets were used. Apart from these segmentations, texture based methods, support vector machine [8], neural networks [9] were used to upgrade the segmentation. But the entire existing works need skull removed image [1, 10].

In contrast to segmentation algorithms, detection algorithms were proposed to deliver tumor existence and output the approximate tumor location instead of providing complete segmentation. For this, unsupervised change detection method used to classify the dissimilar regions across the symmetry line of the brain [11]-[12] and template matching method were proposed [13] recently. According to latest survey [14] many papers focus on segmentation algorithms and not on the image feature extraction. The image feature can explore the tumor types and grades [14]. On account of this, the proposed method developed with the original dataset and with skull, it reduces the segmentation time.

Wavelets are mathematical functions that represent scaled and translated copies of a finite length waveform called the mother wavelet. The wavelets are categorized into continuous and discrete transforms. In discrete transform, the data sets decomposed into high and low pass values repeatedly and called as multi resolution analysis (MRA). The decomposition may take part in all sections is called wavelet packet transformations (WPT). Generally wavelets are used to analyze a signal (image) into different frequency components at different resolution scales. This allows revealing images spatial and frequency attributes simultaneously. The high frequency attributes are used as feature set in this proposed work. The high frequencies derived in various scales are obtained by WPT. From this collection, detection of singular points gives a chance to get edges in the proposed work. Hence modulus maxima are derived to obtain edges [15]-[16]. So the proposed method can take part in boundary based and region based segmentation.

The remaining part of this paper consists of section 2, 3, 4 and 5. In section 2, the basics of wavelet transformation and the proposed framework based on it are described, in section 3, evaluation of significance parameter is given, the results are discussed in section 4 and the conclusion is given in section 5.

2 Method

2.1 Wavelet

The wavelet transforms are classified as continuous wavelet and discrete. The discrete transform works on discrete data. The continuous transform uses all possible scaling factors, starting at 1 and increasing to the number of samples in

the signal. However the continuous wavelet is computationally expensive and for most applications, a dyadic method is used instead. Dyadic wavelet transform is discrete wavelet transform and used only scales that are powers of 2. At the scale s=1, the image is smoothed by convolving it with a smoothing function is stretch and the image is convolved with it again. The process is repeated for s=4, s=8 etc. Since the image is smoothed at each step by a filter bank, the image only contains half of the frequency information and needs half as many samples. So the number of samples in the image is reduced at each stage as well.

The wavelet packet decomposition extends the discrete wavelet transform in a way that each decomposition consists of 2n boxes, generated by a tree of low pass and high pass operations. For n level decomposition there are n+1 possible ways to decompose or encode the signal. The wavelet packet decomposition is shown in Fig.1 .

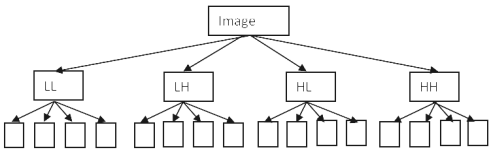


Fig. 1. Wavelet packet transform

The wavelet is represented as a function, where i is the modulation parameter, j is the dilation parameter and k is the translation parameter.

$$\Psi^i_{j,k}(t) = 2^{\frac{j}{2}}\Psi^i(2^{-j}(t-k)) \tag{1}$$

where i=1,2,.. jn and n is the level of decomposition in wavelet packet tree. The wavelet is obtained by the following recursive relationships:

$$\Psi^{2i}(t) = \frac{1}{\sqrt{2}}\sum_{k=-\infty}^{\infty} h_0(k)\psi^i(\frac{t}{2}-k) \tag{2}$$

$$\Psi^{2i+1}(t) = \frac{1}{\sqrt{2}}\sum_{k=-\infty}^{\infty} h_1(k)\psi^i(\frac{t}{2}-k) \tag{3}$$

where h0(k) and h1(k) are discrete filters associated with scaling function and the mother wavelet function. WPT can extract more features [16], so it takes part in the proposed method.

2.2 Proposed Method

The proposed method has the following steps as shown in Fig.2. Initially decompose the image by the implementation of wavelet packet decomposition, in

step 2, the level2 images are composed as mentioned in Fig.2, the modulus maximum image construction is done in step 3, in step 4, mask will be prepared by a thresholding technique and in step 5, the tumor area will be located by the mask.

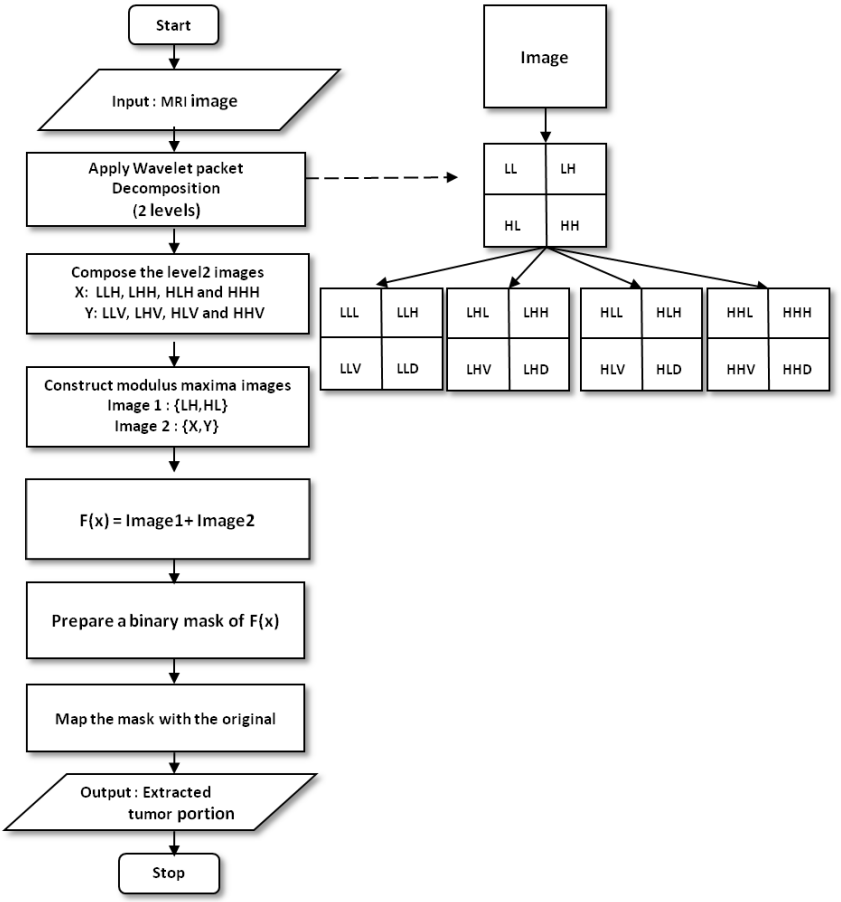


Fig. 2. Flow chart of proposed method

Image Decomposition or Analysis: In this work image feature were extracted by applying (4) and (5) recursively up to 2 levels. The decomposed wavelet packets are shown in Fig. 3.

$$W_{\varphi}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{M-1}^{x=0} \sum_{N-1}^{y=0} f(x, y) \varphi_{j_0, m, n}(x, y) \tag{4}$$

$$W_i^{\psi}(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{M-1}^{x=0} \sum_{N-1}^{y=0} f(x,y)\varphi^i_{j_0,m,n}(x,y) \tag{5}$$

where i=Horizontal, Vertical and Diagonal details.The scaling and translations are defined as (6) and (7)

$$\varphi_{j,m,n}(x,y) = 2^{\frac{j}{2}}\varphi(2^i x - m, 2^j y - n) \tag{6}$$

$$\psi^i_{j,m,n}(x,y) = 2^{\frac{j}{2}}\psi^i(2^i x - m, 2^j y - n), i = H,V,D \tag{7}$$

where j represents scale value, i is the shifting parameter, m and n are the number of rows and columns in a sample.

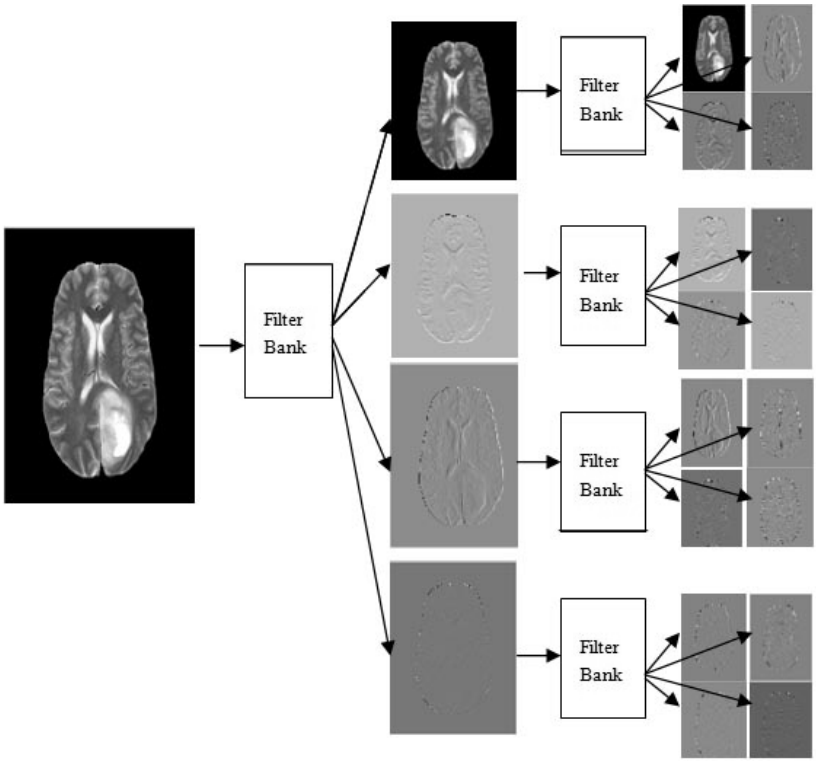


Fig. 3. Wavelet packet decomposition of an image

Image Composition or Synthesis: An image passed through low pass and high pass filters in wavelet filter bank have horizontal edge details and an image passed through high pass and low pass filters have vertical line details. So the composition of all horizontal detail images in level2 from all sections reveal more horizontal edge features and vertical detail sections reveal vertical line features as

shown in Fig.4a and 4b. The following equation is used to compose or synthesize the level2 images.

$$f(x,y) = \frac{1}{\sqrt{MN}} \sum_{m} \sum_{n} W_{\varphi}(j_0, m, n) \varphi_{j_0,m,n}(x,y) \tag{8}$$
$$+ \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_{m} \sum_{n} W_{\psi}^{i}(j,m,n)$$

The proposed method considers Fig.4a and Fig.4b as X and Y.

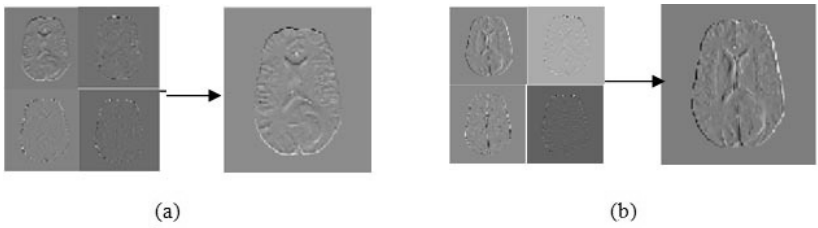


Fig. 4. (a) Composition of LLH, LHH, HLH, HHH (b) Composition of LLV, LHV, HLV, HHV

Modulus Maxima. Already stated that modulus maximum in wavelet is used to extract the singularities and discontinuities. A singularity is surrounded by low pass details [18]. So the modulus maxima images have singularities and isolated low pass details like tumors. The modulus maxima image is obtained by (9)

$$M_s f(x,y) = \sqrt{|W_s^1 f(x,y)|^2 + |W_s^2 f(x,y)|^2} \tag{9}$$

Where f(x, y) is the image value $W_s^1 f(x,y)^2$ and $W_s^2 f(x,y)^2$ are wavelet coefficients of LH and HL parts respectively.

This method uses (9) to obtain two modulus maxima images from level2 and level1. Here it considers the composition of all HL parts in level2 shown in Fig.4a as X and the composition of LH parts shown in Fig.4b as Y. The second one is obtained by using HL and LH parts of level 1 images as x and y. The obtained images are shown in Fig. 5a and Fig. 5b. Then combine these images by adding themselves. The resultant image have isolated tumor area and marked by a square in Fig.5c. This modulus maxima image may lead to segment tumor in either edge based or region based. For region based segmentation the proposed work follows the next step.

Binary Mask. Thresholding is a statistical decision theory to classify the pixels into two or more groups. The isolated tumor area will be extracted by a mask with the help of thresholding and binarization. This method uses the Otsu thresholding which is optimal [19] to prepare a mask. The mask of above image

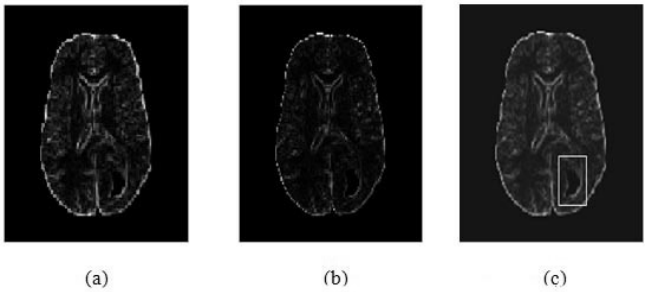


Fig. 5. (a) Composition of LLH, LHH, HLH, HHH (b) Composition of LLV, LHV, HLV, HHV

is shown in Fig.6a. Then the proposed method maps the mask with the original image by replacing the ones and zeros in the mask by zero and the original intensity value respectively is shown in Fig.6b. By intersecting the adjacent slices the exact seed of the tumor will be detected.

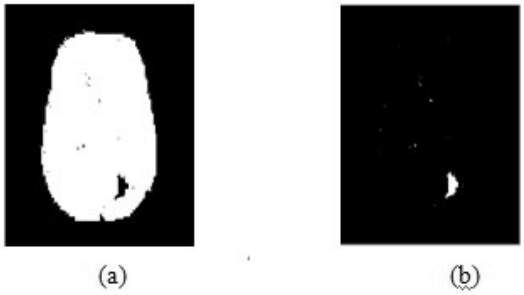


Fig. 6. (a) Mask, (b) Extracted tumor

3 Goodness of Fit

If the proportions of expectation and observation are not equal Chi Square test will be used to identify the goodness of fit. It uses a table to verify the significance of result.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{10}$$

In the above equation O_i represent the observed frequency and E_i represents the expected frequency. In the proposed method, the number of output images which are having tumor in a volume is considered as O_i and total number of slices which are having tumor originally is consider as E_i.

4 Results and Discussion

The experiments are carried out in Matlab7.8 by applying the proposed method on the tumor slices and results are shown in Fig.7. The experiments were done in images with skull and without skull of T1 and T2 weighted images. In Fig.7 odd columns are having the original images the even column are having the extracted images. By this visual result we can conclude that the proposed method works effectively in MRI images with skull and without skull of T1 and T2 weighted images.

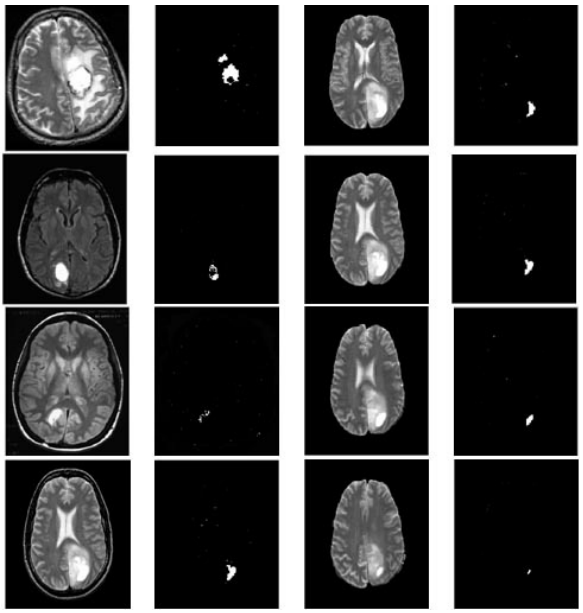


Fig. 7. Samples and output, the odd columns have the original images with skull and without skull, the even columns have the extracted tumor

The goodness of fit is calculated by Chi square test and the results are listed in Table.1. The difference data volumes are collected from scan centres and websites, the data sets varied in slice thickness, tumor quantity and scanning parameter. The random volume consists of sample slices selected from various sources. In Table 1, the MRI brain volume modalities, orientations and skull presence are given in column 2,3 and 4. The total number of slices present in each volume is given in column 5. The tumorous slices identified by medical experts are given as "'Expected"' in column 6. The tumorous slices identified by the proposed method are given as "'Observed"' in column 7. The probability is identified by comparing the total value with the existing chi-square value. According to the Chi square table, the significance (goodness of fit) probability is above 0.75, hence it is noteworthy. The tumor region extracted by the proposed

Table 1. Level of significance

Volume	Modalities	Orientation	Skull	Total Slices	Expected	Observed	χ^2 value
1	T2	Axial	No	28	8	4	2.0
2	T2	Axial	No	55	25	20	1.0
3	T2	Axial	Yes	26	6	3	1.5
4	T2	Axial	Yes	23	9	9	0
5	T2	Axial	Yes	20	4	4	0
6	T2	Coronal	Yes	20	8	8	0
7	T1	Axial	Yes	20	3	4	0.25
8	T2	Axial	Yes	20	4	4	0
9	T2	Axial	Yes	20	7	7	0
10	T1	Axial	Yes	20	5	5	0
11	T2	Axial	Yes	20	11	12	0.0833
12	T1	Axial	Yes	60	5	11	3.2727
Random	All	All	Yes	5	5	4	0.2
Total							**8.03060**

method could be used to quantify the final tumor volume from the MRI images. This work is planned to extend in future.

5 Conclusion

The proposed method need not require any preprocessing algorithm like skull stripping and user intervention for seed selection. It fully depends on the wavelet filters and thresholding. It can locate tumor area from T1 and T2 images and it can assist edge based segmentation and region based segmentation. The slice thickness above 2mm is feasible. It cannot identify the initial stage tumors but if the tumor detected in one slice could be mapped in adjacent slices. This helps to extract the tumor in whole volume. By adding additional technology, the complete tumor extraction and quantization will be done in future.

References

1. Somasundaram, K., Kalaiselvi, T.: Fully automatic brain extraction algorithm for axial T2-Weighted magnetic resonance images. Computers in Biology and Medicine 40, 811–822 (2010)
2. Kalaiselvi, T., Somasundaram, K.: Fully Automatic Method to Identify Abnormal MRI Head Scans using Fuzzy Classification and Fuzzy Symmetric Measure. International Journal on Graphics Vision and Image Processing 10(3), 1–9 (2010)
3. Fletcher Health, L.M., Hall, L.O., Goldgof, D.B., Murtagh, F.R.: Automatic segmentation of non-enhancing brain tumors in magnetic resonance images. Artificial Intelligence in Medicine 21, 43–63
4. Ho, S., Bullitt, E., Gerig, G.: Level-set evolution with region competition: automatic 3-D segmentation of brain tumors. In: Proceeding of International Conference on Pattern Recognition, pp. 532–535 (2002)

5. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Multispectral brain tumor segmentation based on histogram model adaptation. In: Proceeding of SPIE, vol. 6514 65140(5) (2007)
6. Wang, T., Cheng, I., Basu, A.: Fluid vector flow and application in brain tumor segmentation. IEEE Transaction on Biomedical Engineering 56, 781–789 (2009)
7. Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., Ahuja, C.K.: A novel content based active contour model for brain tumor segmentation. Magnetic Resonance Imaging 30, 694–715 (2012)
8. Schoelkopf, B., Smola, A.: Learning with kernels Spport vector machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2002)
9. Jensen, T.R., Schmainda, K.M.: Computer aided detection of brain tumor invasion using multiparametric MRI. J. Magnetic Resonance Imaging 30, 481–489 (2009)
10. Kalaiselvi, T., Somasundaram, K., Vijayalakshmi, S.: A novel self initiating brain tumor boundary detection for MRI. In: Balasubramaniam, P., Uthayakumar, R. (eds.) ICMMSC 2012. CCIS, vol. 283, pp. 464–470. Springer, Heidelberg (2012)
11. Saha, B.N., Ray, N., Greiner, R., Murtha, A., Zang, H.: Quick detection of brain tumors and edemas: a bounding box method using symmetry. Computer Medical Imaging and Graphics 36, 95–107 (2011)
12. Somasundaram, K., Kalaiselvi, T.: Automatic detection of brain tumor from MRI scans using maxima transform. In: National Conference on Image Processing (NCIMP) (2010)
13. Ambrosini, R.D., Wang, P., ODell, W.G.: Computer aided detection of metastatic brain tumors using automated three dimensional template matching. Journal of Magnetic Resonance Imaging 31, 85–93 (2010)
14. Bauer, S., Wiest, R., Nolte, L.-P., Reyes, M.: A survey of MRI based medical image analysis for brain tumor studies. Physics in Medicine and Biology 58, 97–129 (2013)
15. Tu, C.-L., Hwang, W.-L.: Analysis of singularities from Modulus Maxima of Complex wavelets. IEEE Transaction on Information Theory 51 (2005)
16. Mallat: Characterization of signals from Multiscale edges. IEEE Transactions Pattern Analysis and Machine Intelligence PAM1-14, 710–732 (1992)
17. Yang, B., Liu, L., Zan, P., Lu, W.: Wavelet Packet-Based Feature Extraction for Brain-Computer Interfaces. In: Li, K., Jia, L., Sun, X., Fei, M., Irwin, G.W. (eds.) LSMS 2010 and ICSEE 2010. LNCS, vol. 6330, pp. 19–26. Springer, Heidelberg (2010)
18. Bouyahia, S., Mbainaibeye, J.M., Ellouze, N.: Wavelet Based Microclacifications detec-tion in Digitized Mammograms. ICGST GMP Journal 8 (2009)
19. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn., p. 764. Pearson Prentice Hall (2009)

Segmentation of Microcalcifications in Digital Mammogram Images Using Intensity Modified BlackTop-Hat Transformation and Gauss Distribution

P. Shanmugavadivu and S.G. Lakshmi Narayanan

Department of Computer Science and Applications,
Gandhigram Rural Institute - Deemed University, Gandhigram - 624 302,
Tamil Nadu, India
{psvadivu67,sg.lakshminarayanan}@gmail.com

Abstract. An intensity modification based segmentation of microcalcifications from digital mammogram is presented in this paper. The proposed technique projects a novel enhancement method for mammogram images using BlackTop-Hat Transformation and Gauss Distribution as thresholding determinants, taking the neighbouring pixels into consideration for image segmentation. Further, the results are validated with MIAS database description and proved to produce the exact results complying with the descriptions given in the MIAS.

Keywords: Top-Hat Transformation, Gauss distribution, Digital Mammogram, Image Segmentation, Microcalcifications.

1 Introduction

Digital Image Processing (DIP) is a broader processing discipline that processes a digital image subjectively and / or objectively, whereas a digital image is generally defined as an array of discrete intensity values each having unique two-dimensional spatial coordinates [1]. The primary enthusiasm of DIP is to acquire, application-specific information from the images associated with Law enforcement, Defense, Industries, Satellite imaging, Medical and the like. The image processing operations are accomplished through sophisticated algorithms aiming at better visualization and analysis.

Medical image processing focuses on the analysis of the scanned image of the human body. This analysis is used to diagnose the symptoms or to analyse the severity of the diseases by the physicians. The medical imaging modalities are numerous, including radiology, atomic energy, radio waves ultrasound and thermography [2,3]. Breast cancer is highly prevalent among women across the globe. It is very difficult to understand the spreading pattern of diseased tissues as compared to other cancerous cells. The prognosis of breast cancer can easily be accomplished by the digital mammogram images [4,5]. The computer-assisted diagnosis may reduce the complexity of identifying the cancerous cells in the

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 234–240, 2013.
© Springer International Publishing Switzerland 2013

mammogram images [4,5]. For the identification of microcalcifications or other particles computer-aided image analysis is widely used.

Image enhancement, image restoration, image compression and image segmentation are the preprocessing techniques which are commonly performed objectively, prior to image analysis [8-10]. These processes transform the mammogram images more suitable for subsequent processes while identifying the masses in mammogram images. Image segmentation divides the image into small groups of unique pattern or characteristics present in the image.

Morphological operations assist in extracting the objects of an image or describing the image by means of its shape, skeletons, outer region boundaries or edges. Top-Hat transform is used to extract the entire image into smaller regions [7]. The transformation is based on a set of mathematical concepts. In most cases, this approach separates both valleys and peaks present in the transformation using a threshold value. In this article, section 2 describes the methodology to segment the image. Section 3 focuses on the results and discussion. The conclusion is given in the final section.

2 Methodology

Image enhancement, provides better visualization of objects without affecting the original content of the image. In this method the original image is enhanced using morphological operations, which facilitates the identification of boundaries at ease. The Top-Hat transform technique is used for this purpose [6]. The Top-Hat transform is classified into two different operations as WhiteTop-Hat transform and BlackTop-Hat transform. The former one defines an opening operation by using a structural element, whereas the latter uses closing operation with a structural element. In this paper, the BlackTop-Hat Transform (BHT), is implemented using the closing operation. This operation extracts the valley present in the enhanced images. The BHT is obtained by the formula

$$BHT(f) = f \bullet b - f \tag{1}$$

where f is the Euclidean space or discrete grid of the enhanced image, b is the structuring element applied on the gray values of the image, and '$\bullet$'is the closing operation. This BHT operation extracts the enhanced image elements, whose intensity values are smaller than the structuring element. The closing operation can be performed by using the erosion and dilation operations which is represented as:

$$x \bullet y = (x \oplus y) \ominus y \tag{2}$$

The methodology splits the entire proposed work into two phases. Phase I enhances the original image and the second phase of the process applies the Gauss Normalization on the original image, using the formula:

$$f(x) = \frac{1}{\sigma} \times \frac{1}{2\pi} \times e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2} \tag{3}$$

where, μ is the mean of the image which can be derived from the formula $\Sigma \frac{x_i}{N}$ where x_i are pixel values and N is the number of elements in the image, and σ is a variance of the image which can be derived from the formula $\Sigma \frac{(x_i - \overline{x})^2}{N}$ where x_i are pixel values, $\overline{x}$ is the mean of the image and N is the number of elements in the image. The enhanced image is used to locate the brighter spots in the original image, as it increases the intensity of each pixel value. This may produce a better visual effect on the original image. The Gauss distribution is used to eliminate the unwanted information from the original image, which is highly useful while segmenting the required microcalcifications using the enhanced image, as the reference image. The block diagram of the above said methodology is given in Fig.1.

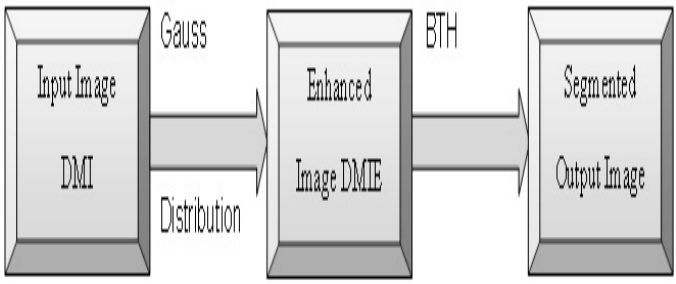


Fig. 1. Block Diagram of the Proposed Method

A. Algorithm for Image Enhancement

Input : A Digital Mammogram (DMI)
Output : Enhanced Image (DMIE)

1. *Read the input image DMI of size* $M \times N$
2. *For the entire image do*
   - *i. Read L(x,y) /* L(x,y)* ε *DMI */*
   - *ii. Compute:*
     - *a. M = Mean (DMImean)*
     - *b. V = Variance(DMIvar)*
     - *c. GD = Gauss Distribution(DMIgd)*
   - *iii. Apply the erosion(x) and dilation(y) on the image as*
     $x \bullet y = (x \oplus y) \ominus y$
   - *iv. Extract the microcalcifications using BHT*
     $BHT(f) = f \bullet b - f$

   where f is a Euclidean space of $DMIE$ *and b, the structuring element. Find the row maximum* $(DMIRmax)$ *and row minimum* $(DMIRmin)$ *intensities for each row of the image*
   - *v. if L(x,y) > DMIgd and L(x,y) < DMIRmax then*
     L(x,y) = L(x,y)+ DMIgd + DMImean
     else
     L(x,y) = DMIRmin
3. *Stop*

The above algorithm enhances the original mammogram image (DMI) into an enhanced image (DMIE) without affecting its original features. From the enhanced image, the microcalcifications in the image are segmented. This algorithm is implemented on the enhanced image thereby the microcalcifications in the image are segmented, using the principle of BHT, as described below.

B. Algorithm for Image Segmentation

Input : Enhanced Image DMIE and original Input image DMI
Output : The segmented image DMI

1. *Read the enhanced input image DMIE*
2. *For the entire image DMIE do*
   i. Read j(x,y) / j(x,y) ε DMIE */*
   ii. if j(x,y) > (DMIEVa + DMIEgd)
   j(x,y)=i(x,y) / i(x,y) ε DMI */*
   else
   j(x,y)=0
3. *For the entire image DMIE*
   i. Read pixels of DMIE, rowwise
   ii. If j(x,y) $\neq$ 0 $\forall$ j(x,y) ε N_4 (j(x,y)) then
   / N_4: Four Neighbourhood */*
   If j(x,y)=0 $\neq$ j(x,y)ε N_8 (j(x,y))
   / N_8: Eight Neighbourhood */*
   iii.Map the boundary of DMIE and DMI.
   iv.Extract the regions from the original image DMI
4. *Stop.*

3 Results and Discussion

This algorithm is implemented in MatLab 7.8. The algorithm was tested on about fifty images with varying units of chaos, angle and magnitude, chosen from Mammogram Image Analysis Society (MIAS) database [14]. It is evident that this method extracts all the microcalcification regions precisely. This computationally simple method has the merit of faster convergence at the designated regions. Using those regions segmented in the enhanced image as the reference, the respective regions are extracted from the original image by mapping the coordinates of the enhanced image on to the original ones.

The obtained results are validated by comparing them with the ground reality descriptions provided by the MIAS dataset. In the MIAS description the first column denotes the feature of background tissue as F - Fatty, G - Fatty-glandular and D - Dense-glandular. The second column specifies the class of abnormalities present in the image as CALC - Calcification, CIRC - Well-defined/circumscribed masses, SPIC - Spiculated masses, MISC - Other ill-defined masses, ARCH - Architectural distortion, ASYM - Asymmetry and NORM - Normal. The third column specifies the severity of the abnormality as B - Benign and M - Malignant. The fourth and fifth column present the co-ordinate pair of centre of

abnormalities and the sixth column denotes the approximate radius of a circle (in terms of pixels) of the abnormality [14].

The initial and final co-ordinate points of the experimental results are obtained by sampling the input images in Adobe Photoshop CS3 version. The comparative description between MIAS specifications and the results of the proposed technique is furnished in Table 1.

Table 1. Spatial Comparision of Mias Datasets and the Obtained Results

Image Reference	MIAS Description	Co-ordinates of Segmented Region (using proposed technique)	
		Initial Position	Final Position
mdb072	F ASYM 725 689 29	673 639	766 712
mdb117	G MISC B 667 365 31	607 311	698 412
mdb148	D MISC M 318 359 27	291 316	348 391

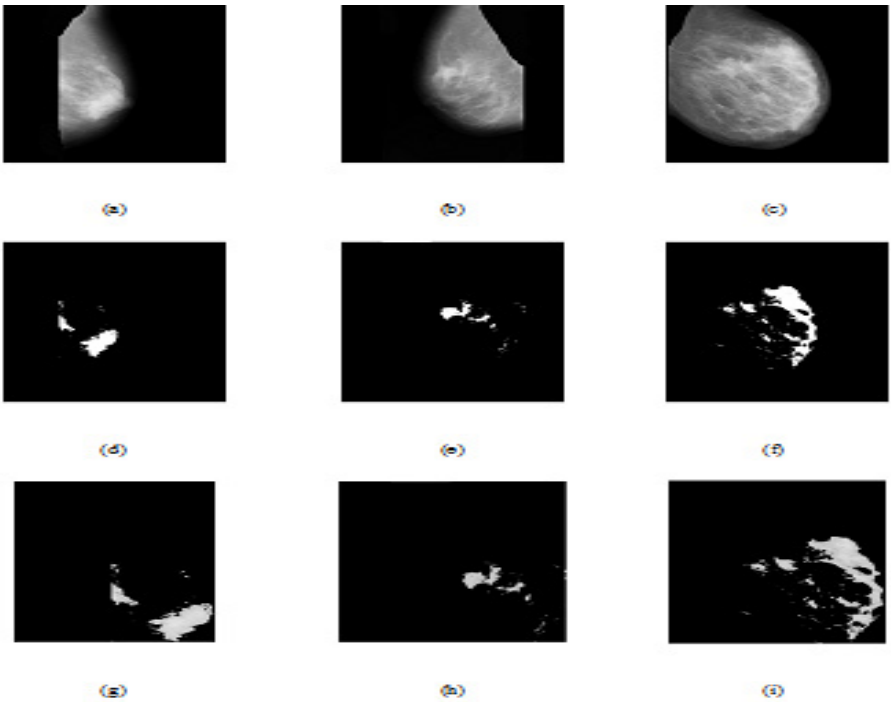


Fig. 2. (a) - (c): Original digital mammogram images (mdb072, mdb117 and mdb148) (d) - (f): Enhanced images of (a) - (c) (g) - (i): Segmented microcalcification regions of (a) - (c)

To illustrate the performance of this technique, three different mammogram images (mdb072, mdb117 and mdb148) were chosen as depicted in Fig.2 (a) - (c) and the subsequent processes performed on them namely image enhancement and segmentation as per the mechanism of the proposed technique are shown in Fig.2 (d) - (f) and Fig.2 (g) - (i) respectively.

The enhancement applied on those images and the corresponding segmented regions in Fig.2 clearly indicate the merit of the proposed method.

It is evident that the proposed technique is sensitive even on small regions and its results are accurate with respect to the ground reality specified by MIAS. This technique hence, segments the suspicious regions of microcalcifications, which are to be further screened through texture analysis.

4 Conclusion

An intuitive segmentation technique presented in this paper has two phases namely image enhancement and image segmentation. The Top-Hat Transformation and Gauss distribution are used to find the optimum threshold values for image enhancement. This technique is computationally simple and is found to accurately segment the microcalcifications regions, irrespective of their size and position. The outcome of this paper provides wider scope for intensity classification for texture analysis.

The input image is enhanced in terms of contrast and brightness enhancement, in order to bring out the intrinsic details, which play a significant role in segmentation. The optimized threshold values computed for BTH, are proved to be highly effective in segmenting the suspicious regions, in a given mammogram. It is apparent from the results that this technique will find a place in the preprocessing phase of texture classification/ analysis of digital mammograms.

References

1. Rafael Gonzalez, C., Richard Woods, E.: Digital Image Processing, 2nd edn. Pearson Education, New Delhi (2002)
2. Shanmugavadivu, P., Lakshmi Narayanan, S.G.: Behavioural Analysis of FCM Clustering on Mammogram Images. In: Proceedings of the National Conference on Signal and Image Processing, pp. 175–177 (2012)
3. Shanmugavadivu, P., Lakshmi Narayanan, S.G.: Segmentation of Microcalcifications in Mammogram Images using Intensity - Directed Region Growing. In: Proceedings of the International Conference on Computer Communication and Informatics ICCCI 2013 (2013)
4. Shanmugavadivu, P., Lakshmi Narayanan, S.G.: Detection of Microcalcifications in Mammogram using Statistical Measures based Region Growing. In: SPIE Proceedings of the International Conference on Communication and Electronics System Design ICESD 2013, vol. 8760, pp. 87601M–1–6 (2013)
5. Shanmugavadivu, P., Lakshmi Narayanan, S.G.: Segmentation of Microcalcification Regions in Digital Mammograms using Self-Guided Region Growing. In: Proceedings of the International Conference on Emerging Trends in Science Engineering and Technology INCOSET 2012, pp. 274–279 (2012)

6. Nevine, H., Eltonsy, G.D., Tourassi, A.S.: A Concentric Morphology Model for the Detection of Masses in Mammography. IEEE Transactions on Medical Imaging 26, 880–889 (2007)
7. Singh, S., Bovis, K.: An Evaluation of Contrast Enhancement Techniques for Mammographic Breast Masses. IEEE Transactions on Information Technology in Biomedicine 9, 109–119 (2005)
8. Cahoon, T.C., Sutton, M.A., Bezdek, J.C.: Breast Cancer Detection using Image Processing Techniques 2, 973–976 (2000)
9. Pfisterer, R., Aghdasi, F.: Detection of Masses in Digitised Mammograms. In: Proceedings of the South African Symposium on Communications and Signal Processing, pp. 115–120. IEEE Digital Library (1998)
10. Chandrasekhar, R., Attikiouzel, Y.: New range-based neighbourhood operator for extracting edge and texture information from mammograms for subsequent image segmentation and analysis. In: IEE Proceedings - Science, Measurement and Technology, vol. 147, pp. 408–413. IEEE Digital Library (2000)
11. MIAS database, UK

Intensity, Shape and Size Based Detection of Lung Nodules from CT Images

K. Veerakumar[1] and C.G. Ravichandran[2]

[1] Department of Electronics and Communication Engineering,
RatnaVelSubramaniam College of Engineering and Technology Dindigul, Tamilnadu, India
[2] Excel Engineering College, Komarapalayam, Tamilnadu, India

Abstract. Lung cancer has become one of the most widely spreading diseases in the world. Detection of lung nodules is the initial step in lung cancer detection. We propose an idea to locate the lung nodules based on its intensity, shape and size. Lung CT images are used for detecting the lung nodules. Initially, Variant Ant Colony Optimization algorithm is used to detect the edges. Variant ACO algorithm greatly helps to reduce the False Positives. Nodules centers are detected in the edge detected image based on the proposed black circular neighborhood algorithm. The intensity of the lung nodules are classified based on the input image using the positions of the lung nodule center. We use lung intensity identification algorithm. Finally the malignant lung nodules are identified from the input CT image based on three features – intensity, shape and size.

Keywords: Edge detection, Ant Colony Optimization (ACO), Intensity based clustering, Image processing, lung CT images.

1 Introduction

A radiologist does the work of detecting the lung nodules from the scan images. Computer Tomography (CT) scan images are widely used to locate the lung nodules. CT screening enables detection of lung cancer at early stages and helps continuous monitoring at later stages. Computer Aided Detection (CAD) helps radiologists to quickly and accurately estimate the size and growth percentage of the nodules. During successive monitoring CAD systems help to detect nodules growth with the help of previous screening and presence of new nodules at present screening.

Lung nodules are of two types – malignant and benign. Lungs nodules are generally spherical in shape. Malignant nodules are larger when compared to benign nodules. A 5–year prospective experience reveals that the diameter of lung cancer will be 5-50mm with an average of 14.4 mm [1]. Lesions greater than 3cm are usually classified malignant and smaller lesions are classified as benign. Nodules with larger diameter should be given greater care when compared to small diameter nodules.

Lung nodule detection starts with segmentation the lung region from the input image. The initial step segments the lung region separately where the nodules are supposed to be located. Then the segmented region is processed for nodule enhancement. Candidate nodule selection is done to identify the lung nodules. Since large number of false positives (FP) is included in the candidate nodule selection stage, we need to go for FP reduction process. These three steps are included in many of the existing

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 241–250, 2013.
© Springer International Publishing Switzerland 2013

systems. We propose a different approach by initially choosing an edge detection algorithm for segmenting. From the edge detected output, we propose a new methodology for candidate nodule selection and FP reduction.

This paper elaborates about a new concept for identifying malignant lung nodules from CT scan images. The idea proposed helps to identify malignant nodules by classifying them with their intensity, shape and size. The CT images are used as input because we use intensity as one of the features in our approach. For edge detection we apply Variant Ant Colony Optimization (Variant ACO) algorithm [2]. Variant ACO is chosen because it helps to reduce the number of false positives at the earlier stage.

We use intensity, shape and size as our features because lung nodules are classified according to these features as malignant and benign. We concentrate on detecting malignant nodules because malignant nodules are cancerous in nature. The final evaluation results shows that we achieve expected output by highly reducing the number of false positives.

2 Related Work

Our approach concentrates greatly on shape and size of the lung nodules. The shape and size are considered as the features for our lung nodule detection system. The lung nodule diameter helps to take further actions for classifying the nodules. When lung nodules are less than 7mm in diameter, follow-up diagnostic is needed [3]. The nodules which are 8-20 mm in diameter need immediate diagnostic CT. And nodules greater than 20 mm diameter CT, PET or biopsy is made and removal is done as done. Other than shape and size features many other features are also used to identify the nodules. Cavouras et., al. proposed solitary pulmonary nodule discriminating (SPND) system consisting of two phases where the first phase uses 20 features with SPN matrix and the second phase uses Least Squares Distance Measure (LSDM) classifier algorithm upon the 20 features. The features are based on the textual features CT density matrix of SPN, SPN density histogram and co-occurrence matrix of SPN. The system differentiates benign and malignant nodules from the CT images. If additional features such as nodules contour, size and CT density measurements, the system may produce higher accuracy rate. Here in our system, we consider size and intensity to improve the efficiency.

An existing system uses the contextual information of the image [4] to identify the lung nodules. Context is described by means of high-level features based on the spatial relation between static contextual features and dynamic contextual features. Initially local features are extracted and in the second phase, context based feature extraction was made. Both local classifier and contextual classifiers were used to classify identifications in a CT scan image. Contextual classification includes Hard exudates binary classification, Drusen binary classification, Cotton wool spots binary classification and Non-lesion binary classification. When conceptual information is used, the CAD system is similar to human's observation. An expert suggested using the whole image instead of selecting only the specific ROI.

Most lung cancer detection methods involve segmentation in their first step. Segmentation of CT image involves the segmentation of lung itself, the airways and vessels portion and finally the lobar segmentation. Lung cancer detection is done by detection of pulmonary nodules, characterization of pulmonary nodules, nodular size measurement. The detection of nodules is done by primarily selecting the candidates

using one of the following method viz. multiple gray-level thresholding, mathematical morphology, connected component analysis of thresholded images. False positives are reduced in the following process along with the classification. Fast lung nodule detection is proposed [5] which undergo the normal method for lung nodule detection of lung segmentation, nodule enhancement, and finally false positive reduction. For nodule enhancement, cylindrical shape filter is used and Support Vector Machine (SVM) with seven types of parameters is applied to reduce the false positives. It fails to detect nodules that are close to or sticking to the lung walls or blood vessels. As we consider the whole image, this has been overcome.

Adaptive distance based threshold is also used to identify lung nodules. Using this, Fisher Linear Discriminant (FLD) classifier is used. The dataset created in the process was by the Standard Digital Image Database Project Team of the Scientific Committee of the Japanese Society of Radiological Technology (JRST).

A large number of trials are attempted to reduce the number of False Positives (FP) after identifying the lung nodules. Murphy et al., [6] present a system with which uses the local features like shape index and curvature. Here two successive k-nearest-neighbor (KNN) is applied to reduce the false positives. Using the shape index and curvedness, seed points within the lungs are detected by thresholding. A region growing algorithm is applied to form clusters which are then merged and location adjustment is made. Upon the clusters, KNN algorithm is applied. To reduce the false positives, again KNN clustering algorithm is applied and nodules are localized. To reduce the detection of unwanted lung nodules, a vowel-based neural approach is also applied [7]. Here, spherical shape objects are detected and neural approach is used to detect the lung nodules from the CT images. The lung portion is segmented and nodule candidates are selected. This approach identifies the region of interest using ROI hunter algorithm. It eliminates false positives using vowel-based approach. The nodules listed is classified using a neural classifier and tagged as nodule if voxels percentage is higher that a certain threshold. It is helpful for finding nodules whose diameter is greater than 5mm only. An Iris filter is used to discriminate the false positives with the nodules [8]. Regions were characterized by iris filter output and morphological features are extracted from CT images. To reduce the false-positives iris filter is used with the Linear Discriminant Analysis (LDA). As the name suggests, the iris filter can be used as second reader for radiologists.

A study has been made by applying many systems over the same set of images [9]. The dataset selected was ANODE09. Six algorithms viz., Fujitalab, region growing volume plateau, Channeler Ant model, Voxel-based neural approach, ISI-CAD and Philips Lung Nodule CAD are applied and their performance were compared. Finally a system which combines all the approaches was proposed. The detection of 80% nodules suggests that blending of algorithms give a right path for new inventions.

3 Methodology

The detection and analysis of lung nodules is done by locating the nodule in the CT image followed by analysis of located nodules. Three steps are followed in the system.

1. Edge detection.
2. Shape detection.
3. Intensity based classification.

3.1 Edge Detection

Edges are change in intensity of the image. Ant Colony optimization algorithm (ACO) has been suggested for detection of edges from the CT lung images. The ACO algorithm is based on the idea of pheromone left by the ants. The ants following the previous ants use the pheromone left by its predecessor ants. ACO algorithm leaves behind a large number of false positives. To reduce the number of false positives, Variant ACO algorithm can used to detect the edges of the image [2]. Variant ACO algorithm developed is an alteration of Act Colony Optimization algorithm. Variant ACO algorithm works on the output of another edge detected image. The lung CT image is edge detected using the Otsu algorithm. Then, the ACO algorithm is applied which calculates the probability for the next stage using the previous stage's decision. By taking different paths, the algorithm chooses the best path which gives the edge if the image. By using Variant ACO method, noise in the output is very much reduced. False Positives are reduced at initial stage itself. The Variant ACO algorithm is given in algorithm 1.

Algorithm 1: Ant colony optimization algorithm

```
Input: Lung CT image.
Output: Edge detected lung CT image.
begin
    for each pixel of image
        find weight of pixel for both foreground and background
        find mean of weights for both foreground and background
        find variance of the pixels
    end for
    calculate within class variance for all the pixels
    maximum value of within class variance gives the edge pixels.
    Initialize the base attractiveness, τ, and visibility, η, for each edge
    for i < IterationMax do:
        for each ant do:
            choose probabilistically (based on previous equation) the
        next state to move into;
            add that move to the tabu list for each ant
            repeat until each ant completed a solution
        end
        for each ant that completed a solution do:
           update attractiveness τ for each edge that the ant traversed
        end
        if (local best solution better than global solution)
        save local best solution as global solution
    end
end
```

Fig. 1. A Input CT scan image 1b. Output from Otsu algorithm 1c. Output of variant ACO algorithm

3.2 Shape and Size Detection

Lung nodules (both malignant and benign) are mostly spherical in shape. Lung nodules can be classified based on size. Cancerous (malignant) nodules are larger when compared to non – cancerous (benign) nodules. More care should be given to the nodules whose size is greater than 3 cm. While detecting lung nodules, the size has to be detected along with the shape. A spherical region with the specified diameter gives the nodules present in the given CT image.

The edge detected image is given as the input to black circular neighborhood algorithm. The assumption is that all the edges are black pixels. The edges here represents the edges of the lungs also the nodule's edges. As the lung edges are fine lines and nodules are identified in the form of clustered black pixels, we apply black circular neighborhood algorithm. The work of black circular neighborhood algorithm is to find the center pixel of clustered black pixels. Initially, clustered black pixels are identified. The size of the black pixel clusters are used to locate the nodules from the CT image. From the black pixel clusters, the center pixel is identified to be the center of lung nodules. Black circular neighborhood algorithm is presented in algorithm 2.

Using the spherical shape and size we have spotted the nodules in the edge detected image. These nodules have to be classified based on their intensity. The center of lung nodules are given to the intensity based lung nodule identifier algorithm.

Algorithm 2: Black circular neighborhood algorithm

```
  Input: Variant ACO algorithm output image.
  Output: Pixels of nodules center.
  Assumption: Variant ACO output is B/W image with Black
pixels representing edges.
  begin
       detect edge pixel positions
       for each edge pixel
          check for 8-neighborhood
          if accepted add to center_pixels
       end for
       return center_pixels
  end
```

3.3 Intensity Based Clustering

Nodules are identified by the radiologists with the help of intensity change in the CT image. A whiter region on the image indicates the presence of nodule in lungs. We use this intensity feature of the nodule to identify the nodules in our system. In this paper we propose an intensity identification algorithm for locating the nodules position over the CT image.

The center pixels of the lung nodules are retrieved from the black circular neighborhood algorithm. After identifying the nodules center pixel, the intensity is used to classify the nodule for detecting malignant nodules. A circular region is determined around the center pixel of the nodules. The intensity of the nodules within the radius of the circular region is retrieved from the input lung CT image. The sum of intensity of all the nodules inside the circular region is calculated and its average value is found. When the average intensity exceeds the threshold intensity, the specified pixel is marked as nodule pixel and color is modified to indicate the nodule.

Algorithm 3: Intensity based clustering algorithm

```
Input: Nodules center pixels, Input CT image.
Output: Nodule pixels.
begin
    for each center_pixel
        find pixels within circular radius
        find pixels intensity
        determine circular region with figured pixels
        calculate the average intensities within the circular region
        if avg>threshold
            assign center_pixel to nodule_pixel
            for each center_pixel
                modify pixel color of nodule
            end for
        end if
    show nodule highlighted image
end
```

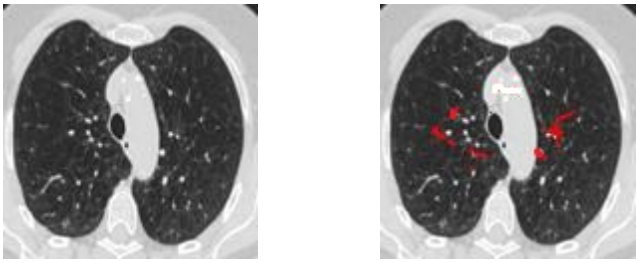


Fig. 2. A. Input CT image with 15 nodules b. Nodules (10 True positives) detected using Intensity based clustering technique.

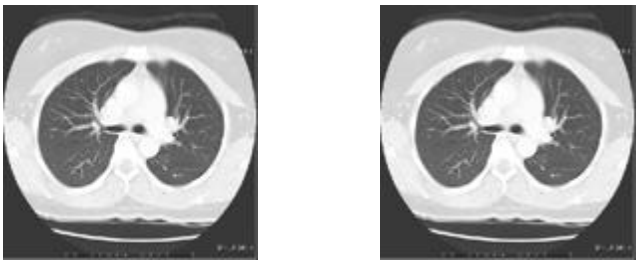


Fig. 3. A Input CT image with no nodules b. Output with no detection of nodules.

4 Results and Discussion

The proposed idea is evaluated for a dataset with 33 images. We use accuracy and precision to check the accuracy of the proposed algorithms. True positive gives the number of lung nodules correctly identified true negatives indicates unidentified nodules. The number of wrongly identified nodules is termed as false positive and the correctly rejected portions are false negatives.

$$\text{Accuracy} = \frac{\text{Number of true positive + number of true negative}}{\text{No. of true positive + false positive + false positive + true negative}} \quad (1)$$

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{Number of true positive + false positives}} \quad (2)$$

Accuracy is measured for 33 different CT lung images. The nodules which has the diameter more than 4 units and whose intensity is higher than 200 (we have used gray scale image) is taken as malignant nodule in our study. Data value for all the 33 images is shown in fig. 4. The graph showing the accuracy values for all the images is shown in fig.5. From the figure, we can see that, the accuracy of the algorithm is above 0.65. When the number of nodules is more in the input image, the accuracy of the system is higher. Similarly, precision is also observed for the input image dataset.The average precision value of the proposed idea is 0.73. Fig. 6 shows the precision graph for 33 images. As the precision parameter is used to evaluate the proposition of correctly identified nodules, we use it to evaluate the proposed algorithm.

Sensitivity measures the proposition of true positives and specificity measures the proposition of true negative.

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives + number of false positives}} \quad (3)$$

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives + number of false negatives}} \quad (4)$$

Image ID.	Sensitivity	Specificity	Accuracy	Precision
1	0.66666667	0.83333333	0.71428571	0.90909091
2	0.5	0.5	0.5	0.33333333
3	0	1	0.66666667	1
4	0.58333333	0.85714286	0.68421053	0.875
5	0.66666667	0.5	0.57142857	0.5
6	0.5	1	0.66666667	1
7	0.5	0.5	0.5	0.5
8	1	1	1	1
9	0.8	0.33333333	0.625	0.66666667
10	1	0	0	0
11	0.72727273	0.75	0.73684211	0.8
12	1	1	1	1
13	0.72727273	0.66666667	0.70588235	0.8
14	0.58333333	0.625	0.6	0.7
15	1	1	1	1
16	0.83333333	0.5	0.7	0.71428571
17	0.55555556	0.5	0.53846154	0.71428571
18	0.6	0.66666667	0.625	0.75
19	1	1	1	1
20	0.5	1	0.66666667	1
21	0.63636364	0.71428571	0.66666667	0.77777778
22	0.66666667	0.66666667	0.66666667	0.66666667
23	0	0.5	0.33333333	0
24	0.5	0.66666667	0.6	0.5
25	1	1	1	1
26	0.63636364	0.71428571	0.66666667	0.77777778
27	0.58333333	0.75	0.65	0.77777778
28	0.25	0.66666667	0.42857143	0.5
29	0.5	0.66666667	0.6	0.5
30	0.66666667	0.83333333	0.73333333	0.85714286
31	1	1	1	1
32	0.69230769	0.71428571	0.7	0.81818182

Fig. 4. Data values for evaluation parameters

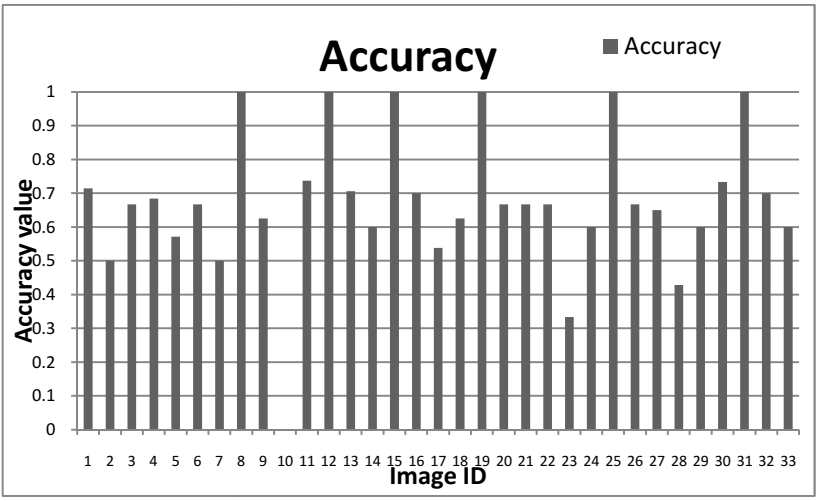


Fig. 5. Accuracy of the proposed intensity based clustering algorithm

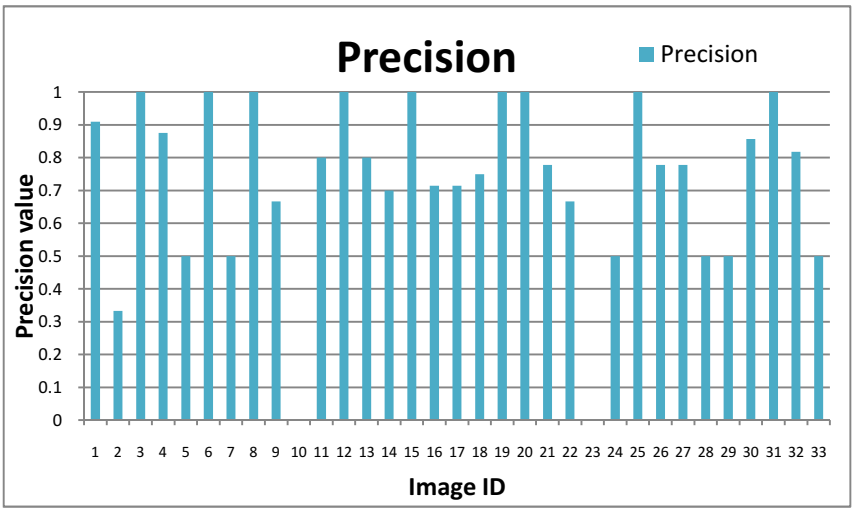


Fig. 6. Precision for 33 images shows the proposition of true positives

Fig 7 shows the sensitivity and specificity percentage for 33 images. Sensitivity has reached 100% for many images. Specificity is also high. Average sensitivity is 0.65 and average specificity is 0.72.

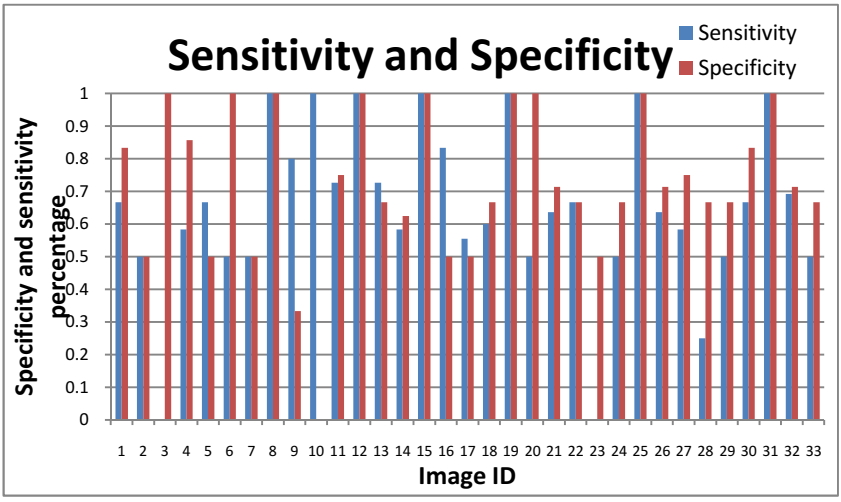


Fig. 7. Sensitivity and specificity for input image dataset (33 lung CT images)

5 Conclusion

The system helps to identify malignant nodules typically based on shape, intensity and size. For reducing the number of false positives (FPs), we use Variant Ant Colony

Optimization algorithm. We can use Iris filter [8] to further reduce the number of FPs. The size based filtering of nodules helps to reduce the clustering process and in clustering, the intensity based clustering identifies the percentage of nodules growth. The system can be extended to detect benign lung nodules using size feature.

The system can be extended to detect nodules other than spherical one, when we use axis values. While expanding the circular region towards either x-axis or y-axis, we can find elliptic nodules and larger nodules. This helps in finding out benign nodules.

References

1. Leung, A., Smithuis, R.: Solitary pulmonary nodule: benign versus malignant Differentiation with CT and PET-CT. Radiology (May 20, 2007), http://www.radiologyassistant.nl/en/p460f9fcd50637
2. Veerakumar, K., Ravichandran, C.G.: Applying Ant Colony Optimization algorithms and variants for lung nodule detection. Pattern Analysis and Applications (2013) (Communicated)
3. Murphy, K., van Ginneken, B., Schilham, A.M., de Hoop, B.J., Gietema, H.A., Prokop, M.: A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Medical Image Analysis 13(5), 757–770 (2009)
4. Retico, A., Delogu, P., Fantacci, M.E., Gori, I., Preite Martinez, A.: Lung nodule detection in low-dose and thin-slice computed tomography. Computers in Biology and Medicine 38(4), 525–534 (2008)
5. Sánchez, C.I., Niemeijer, M., Išgum, I., Dumitrescu, A., Suttorp-Schulten, M.S.A., Abràmoff, M.D., van Ginneken, B.: Contextual computer-aided detection: Improving bright lesion detection in retinal images and coronary calcification identification in CT scans. Medical Image Analysis 16(1), 50–62 (2012)
6. Suárez-Cuenca, J.J., Tahoces, P.G., Souto, M., Lado, M.J., Remy-Jardin, M., Remy, J., Vidal, J.J.: Application of the iris filter for automatic detection of pulmonary nodules on computed tomography images. Computers in Biology and Medicine 39(10), 921–933 (2009)
7. Swensen, S.J., Jett, J.R., Hartman, T.E., Midthun, D.E., Mandrekar, S.J., Hillman, S.L., Sykes, A.-M., Aughenbaugh, G.L., Bungum, A.O., Allen, K.L.: CT screening for lung cancer: five-year prospective experience. Radiology 235(1), 259–265 (2005)
8. Atsushi, T., Fujita, H.: Fast lung nodule detection in chest CT images using cylindrical nodule-enhancement filter. International Journal of Computer Assisted Radiology and Surgery, 1–13 (2013)
9. van Ginneken, B., Armato, S.G., de Hoop, B., van de Vorst, S., Duindam, T., Niemeijer, M., Murphy, K., et al.: Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Medical image analysis 14(6), 707–722 (2010)

Detection of Cardiac Abnormality from Measures Calculated from Segmented Left Ventricle in Ultrasound Videos

G.N. Balaji and T.S. Subashini

{balaji.gnb,rtramsuba}@gmail.com

Abstract. In this paper a novel and robust automatic LV segmentation by measuring the properties of each connected components in the echocardiogram images and a cardiac abnormality detection method based on ejection fraction is proposed. Starting from echocardiogram videos of normal and abnormal hearts, the left ventricle is first segmented using connected component labeling and from the segmented LV region the proposed algorithm is used to calculate the left ventricle diameter. The diameter derived is used to calculate the various LV parameters. In each heart beat or cardiac cycle, the volumetric fraction of blood pumped out of the left ventricle (LV) and the ejection fraction (EF) were calculated based on which the cardiac abnormality is decided. The proposed method gave an accuracy of 93.3% and it can be used as an effective tool to segment left ventricle boundary and for classifying the heart as either normal or abnormal.

Keywords: Echocardiogram, Left ventricle, automatic detection, segmentation, region props, ejection fraction.

1 Introduction

The human heart has four chambers, two superior atria (right atria and left atria) and two inferior ventricles (right ventricle and left ventricle), in which atria are the receiving chambers and ventricles are the pumping chambers. Blood flows from the atria to the ventricles in one directional. The de-oxygenated blood is collected by the right side of the heart from the right atrium and pumps it to lungs through right ventricle where the oxygen is picked up. The oxygenated blood in the lungs is collected by left atrium in the left side of the heart. Blood moves from left atrium to the left ventricle which pumps it out to the body as shown in the block diagram of Fig. 1. Due to improper blood flow heart failure occurs [1]. Heart failure begins mostly in left ventricle which is the main pumping chamber [2]. The confirmation of heart failure is done by echocardiography. To see the heart valves, heart beating and other structures of heart echocardiogram is used. For the optimal solution of the heart systolic and diastolic phases are linked closely in [3], where the systolic is the contraction and diastolic is the relaxation and filling. The contraction and relaxation will provide a good diastolic reserve for a normal heart, whereas for an abnormal heart due to disease process the LV contraction and relaxation will be abnormal. For accurate diagnosis of heart related problems clinical parameters such as ejection

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 251–259, 2013.
© Springer International Publishing Switzerland 2013

fraction (EF), myocardial mass (MM) are the requirements by the cardiologist. The ejection fraction the range must be 55-70% for the normal heart and for the abnormal heart the ejection fraction cutoff will be 40% [15]. Heart failure is mostly due to the main pumping chambers, which may become stiff and not fill properly between the beats. For the evaluation of heart muscle, heart valves and risk for heart diseases echocardiogram is the easiest and widely employed method that uses ultrasound. Though left side, right side or both sides of the heart can involve in heart failure, the heart failure begins with the left side typically - specifically the main pumping chamber left ventricle. In each heart beat or cardiac cycle, the volumetric fraction of blood pumped out of the left ventricle (LV) represents the ejection fraction (EF) and the cardiac abnormality is decided based on ejection fraction. An automatic detection of cardiac abnormality based on ejection fraction calculated from segmented left ventricle in ultrasound video is proposed in this paper.

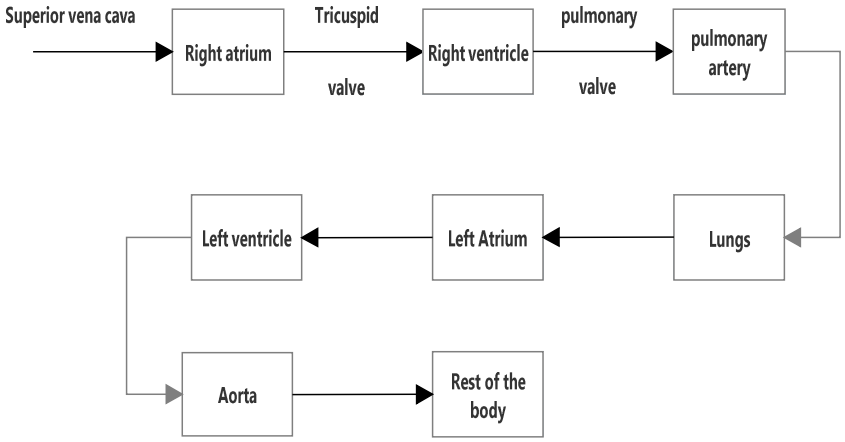


Fig. 1. Blood circulation through the heart

2 Previous Work

Worldwide, Cardiovascular disease such as myocardial ischemia is the leading cause of death. Because of the low contrast and presence of noise the edge detection algorithm fails for reducing the noise without distorting the clinical features [4].For noise suppression in ultrasound images wavelet based thresholding was proposed in [5]. Detecting epicardial and endocardial boundary in short axis echocardiographic sequences using a multiple active contour model in [6]. The automatically-segmented areas show excellent agreement with manually segmented areas, measured by a specialist. The proposed methods could be used to eliminate inter and intra-observer variations that are typically observed in manual border delineation in [7]. To incorporate temporal information into the level set method as a curve evolution regularize to solve the boundary leakage problem caused by dropouts when we detect

the inner heart wall boundary from echocardiographic image sequence a new method is proposed in [8]. To locate the region containing the LV a watershed transform and morphological operation and for the detection of the endocardial boundary of the LV which performs snake deformation with a multi-scale directional edge map in [9]. For segmentation and boundary detection active contour model is combined the K-Means clustering algorithm was proposed in [10]. The location of the LVCP is tracked automatically and the parameters are adjusted adaptively localization of ROI which reduces the interference from artifacts and also provides an initial contour for LV boundary detection was proposed in [11]. For unsupervised computation of ejection fraction from raw echocardiogram videos an automatic method is proposed in [12].

3 Methodology

3.1 Preprocessing

The block diagram of the proposed system is shown in Fig.2. Firstly the frames are extracted from the echocardiogram video, then Noise is removed and contrast is enhanced in each of the frames by applying high boost filter and LoG. This results in the region of interest namely left ventricle to be highlighted.

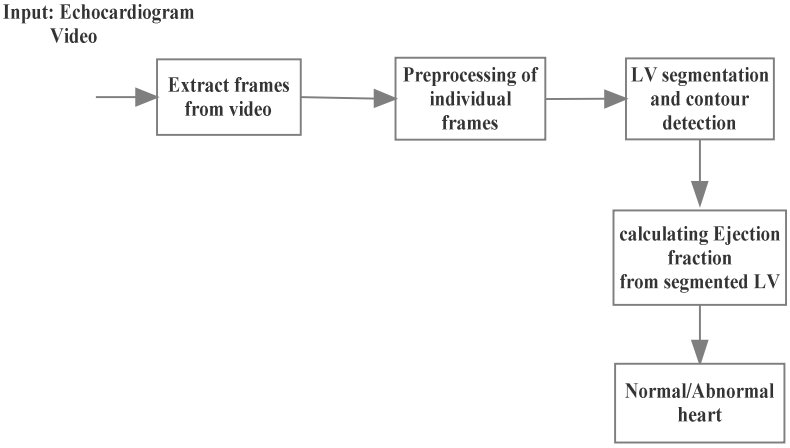


Fig. 2. The proposed system

An extracted frame is shown in Fig. 3(a). It is often desirable to emphasize high frequency components representing the image details without eliminating low frequency components (such as sharpening). The high-boost filter can be used to enhance high frequency component. The high boost filtering is expressed in equation form as follows.

$$\text{Highpass} = \text{original} - \text{lowpass} \tag{1}$$

If original is multiplied by an amplification factor A, a high boost image will result,

$$\text{High boost} = (A)*(\text{original})\text{-lowpass} \tag{2}$$

$$= (A\text{-}1)*(\text{original}) - \text{lowpass} + A*(\text{original} - \text{lowpass}) \tag{3}$$

$$= (A\text{-}1)*(\text{original}) - \text{lowpass} + \text{highpass} \tag{4}$$

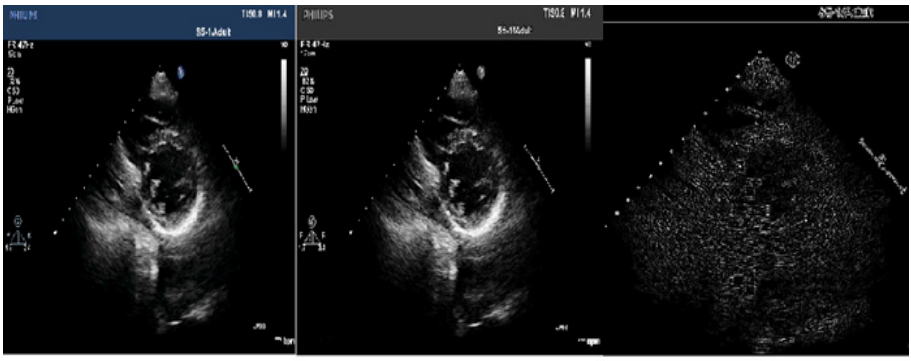


Fig. 3. (a) Extracted frame, (b) after applying high boost filter, (c) LoG applied image

Fig. 3(b) shows the high boost filtered image, it is common to smooth the image (e.g., using a Gaussian filter) before applying Laplacian filters, since Laplacian filters are second order derivative filters. This two-step process is calling the Laplacian of Gaussian (LoG) operation.

First the image is smoothened using Gaussian operator

$$\mathrm{h}(\mathrm{x},\mathrm{y}) = \exp\left(\frac{-x^2-y^2}{2\sigma^2}\right) \tag{5}$$

Where σ represents standard deviation,

Then Laplacian operator is applied

$$x^2 + y^2 = r^2 \tag{6}$$

$$\text{Laplacian } \nabla^2 h = \left(\frac{r^2-\sigma^2}{\sigma^4}\right) exp\left(\frac{-r^2}{2\sigma^2}\right) \tag{7}$$

The LoG finds the areas of rapid change (edges) in images as shown in Fig. 3. (c).The segmentation of LV is the next step. The steps involved in segmentation process are shown in Fig. 4.

3.2 Segmentation

To obtain the cardiac cavity in the processed frame morphological closing is performed which is labeled using connected component labeling as shown in Fig. 5. (a) and 5(b).

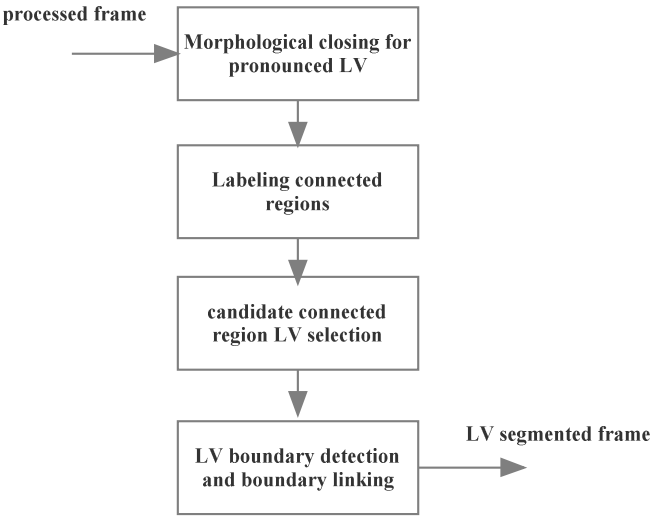


Fig. 4. Steps involved in segmentation of LV

Connected-component labeling works by scanning an image, pixel by pixel (from top to bottom and left to right) in order to identify connected pixel regions, i.e. Regions of adjacent pixels which share the same set of intensity values V [13].

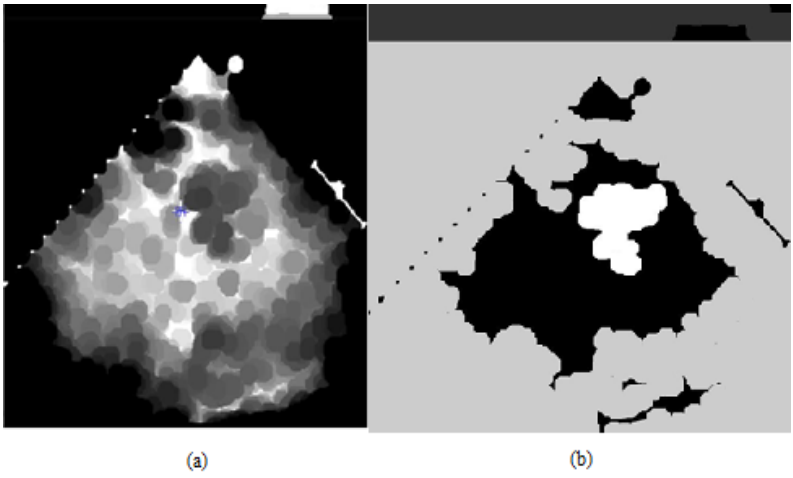


Fig. 5. (a) Morphological closing (b) Connected-Component labeling

The candidate region LV is chosen by discarding the smaller labeled components as shown in Fig. 6(a). After LV detection boundary detection and boundary linking is performed as shown in Fig. 6(b).

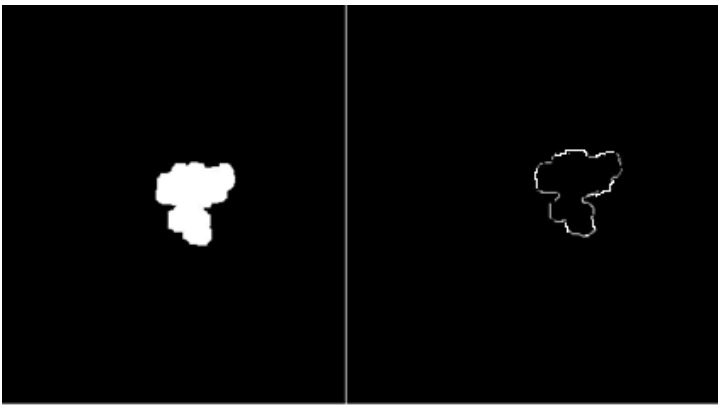


Fig. 6. (a) segmented LV, (b) After boundary linking

3.3 LV Parameters

Using region props equivalent diameter is calculated for all segmented LV. The average radius of the LV is also calculated based on which global parameters like ejection fraction and LV cavity area are extracted. To find the left ventricular volume the following equation is used [4].

$$Volume = \left(\frac{7.0}{2.4 + D}\right) D^3 \tag{8}$$

where 'D' is the average endocardial diameter.

For the detection of cardiac abnormality left ventricle ejection fraction (LVEF) is used. In healthy individuals, the LVEF is typically greater than 0.5 (50%) [15].

The end diastolic volume (EDV) is the largest cavity volume throughout the cardiac cycle, and end systolic volume (ESV) is the smallest cavity volume throughout the cardiac cycle. Stroke volume (SV) which is the difference between EDV and ESV is then divided by the end diastolic volume (EDV) to obtain the ejection fraction.

$$LVEF = \frac{SV}{EDV} \tag{9}$$

$$where\ \ SV = EDV - ESV \tag{10}$$

Where EDV is the end diastolic volume and ESV is the end systolic volume.

4 Results and Discussion

A dataset of 50 patients consisting of 30 normal hearts and 20 abnormal hearts were taken up for this study. The dataset is collected from the Cardiology department of the

Raja Muthaiah Medical College Hospital, Annamalai University, with the help of medical experts taken with the new iE33 xMATRIX echo system. The videos are up to 4 to 6 seconds having around 46 frames per second. The resolution of the frame is 1024 × 768 pixels. Fig. 7 shows the LV volume of a normal and an abnormal heart graphed for one cardiac cycle of the heart. The cardiac cycle of the normal and abnormal heart taken up in this figure goes up to 8 frames. The X-axis shows the frame number and the Y-axis gives the volume of the heart in each frame.

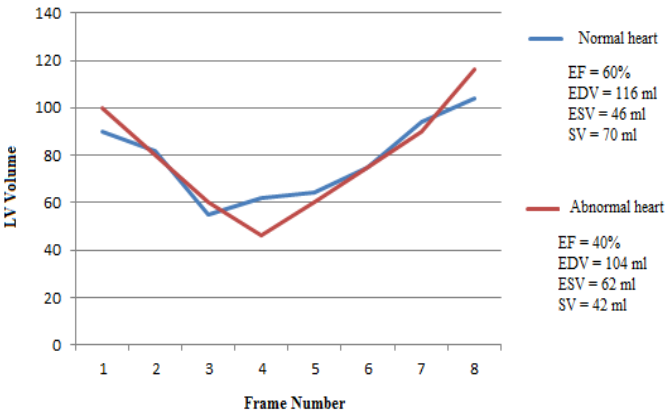


Fig. 7. LV Volume along the echo video

The values shown in the right side of the graph shows the parameters calculated using the left ventricle volumes. Normally in healthy individuals, the left ventricle ejection fraction LVEF is typically greater than 0.5 (50%) [15]. It can be seen from the graph that LVEF is 60% for normal heart and 40% for abnormal heart. Table 1 shows the PSNR value for different denoising techniques applied to echocardiogram images. From Table1 it can be seen that the proposed high boost filter gives high PSNR value. The proposed edge detection algorithm was applied and the accuracy of the results was verified with the cardiologist of the Raja Muthaiah Medical College Hospital, Annamalai University. From Fig. 8(a) and 8(b) it can be seen that the proposed edge detection algorithm is able to detect the LV boundary accurately and it almost coincides with manually sketched LV by the specialist.

The ejection fraction values are used to classify the heart as normal or abnormal. Among the 30 normal hearts in dataset 29 normal hearts are classified correctly and among the 20 abnormal hearts 18 were classified correctly. Table 2, shows the classification of normal and abnormal hearts and the overall accuracy is 93.3%.

Table 1. PSNR values of denoising filters

Denoising method	PSNR	Reference
LP transform based despeckling method	28.44	[16]
WT-L3-ST (bior 6.8)	29.14	[16]
Wiener (3x3)	32.14	[16]
High boost filter	**33.16**	**proposed**

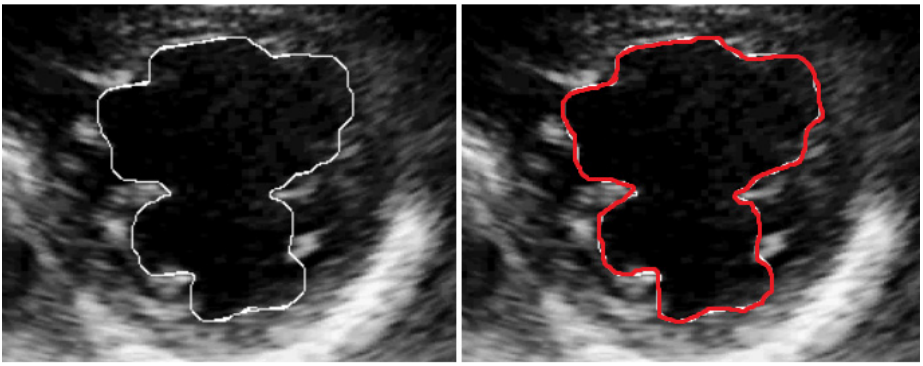


Fig. 8. (a) Automatically segmented LV (b) Manually segmented LV

Table 2. Classification of heart using left ventricle ejection fraction (LVEF)

Type of heart	No videos taken	No of videos classified correctly	Accuracy
Normal	30	29	96.6 %
Abnormal	20	18	90 %
		Overall accuracy	93.3%

5 Conclusion

A novel and robust method for automatic LV segmentation and cardiac abnormality detection based on ejection fraction is proposed in this paper. The ejection fraction values are used to classify the heart as normal or abnormal. The proposed method gave an accuracy of 93.3% and it can be used as an effective tool to segment left ventricle boundary and for classifying the heart as normal or abnormal.

Acknowledgment. The authors would like to thank Dr. E. Balasubramaniyan and Dr.A.Manikandarajan of Raja Muthaiah Medical College Hospital, Annamalai University for their suggestions and for the verification of the results.

References

1. Definition of Heart failure, Medical Dictionary, MedicineNet (April 27, 2011)
2. Heart failure, Health Information, Mayo Clinic (December 23, 2009)
3. Cheng, C., Noda, T., Nozawa, T., Little, W.: Effect of heart failure on the mechanism of exercise induced augmentation of mitral valve flow (1993)
4. Allender, S., Peto, V., Scarborough, P., Kaur, A., Rayner, M.: Coronary heart disease Statistics. British Heart Foundation Statistics Database (2008)
5. Sudha, S., Suresh, G.R., Sukanesh, R.: Speckle Noise Reduction in Ultrasound Images by Wavelet Thresholding based on Weighted Variance. International Journal of Computer Theory and Engineering (2009)
6. Chalana, V., Linker, D.T., Haynor, D.R., Kim, Y.: A multiple active contour model for cardiac boundary detection on echocardiographic sequences. IEEE Transactions on Medical Imaging (1996)
7. Reis, M.D.C.D., da Rocha, A.F., Vasconcelos, D.F., Espinoza, B.L.M., Nascimento, F.A.D.O., Carvalho, J.L.A.D., Salomoni, S., Camapum, J.F.: Semi-Automatic Detection of the Left Ventricular Border. In: 30th Annual International IEEE EMBS Conference, August 20-24 (2008)
8. Fang, W., Chan, K., Fu, S., Krishnan, S.M.: Incorporating Temporal Information Into Level Set Functional for Robust Ventricular Boundary Detection From Echocardiographic Image Sequence. IEEE Transactions on Biomedical Engineering (2008)
9. Jierong, C., Foo, S.W., Krishnan, S.M.: Watershed pre-segmented snake for boundary detection and tracking of left ventricle in echocardiographic images. IEEE Transactions on Information Technology in Biomedicine (2006)
10. Nandagopalan, S., Dhanalakshmi, C., Adiga, B.S., Deepak, N.: Automatic Segmentation and Ventricular Border Detection of 2D Echocardiographic Images Combining K-Means clustering and Active Contour Model (2010)
11. Jierong, C., Foo, S.W., Krishnan, S.A.: Automatic detection of region of interest and center point of left ventricle using watershed segmentation. In: IEEE International Symposium on Circuits and Systems (2005)
12. Beymer, D., et al.: Automatic estimation of left ventricular dysfunction from echocardiogram videos. In: IEEE Conference on Computer Society (2009)
13. Gonzanez, R.C., Woods, R.E.: DIP. Pearson education Singapore (2002)
14. Force, T.L., Folland, T.D., Aebischer, N., Sharma, S., Parisi, A.F.: Echocardiographic Assessment of Ventricular function. Cardiac Imaging (1991)
15. Kumar, V., Abbas, A.K.A., Jon: Robbins and Cotran pathologic basis of disease (8th). Elsevier Saunders, St. Louis (2009)
16. Zhang, F., Koh, L.M., Yoo, Y.M., Kim, Y.: Nonlinear diffusion in Laplacian pyramid domain for ultrasonic speckle reduction. IEEE Trans. on Medical Imaging (2007)

A Comparative Study of Wavelet Coders for Image Compression

PL. Chithra and K. Srividhya

Department of Computer Science, University of Madras, Chennai - 600 005, India
chitrasp2001@yahoo.com, sribhuvan@gmail.com

Abstract. This paper focuses on comparison of different wavelet coders such as SPIHT, SPECK, BISK and TARP for efficient storage and better transmission. Set partition methods like SPIHT, SPECK and BISK (variant of SPECK) are based on the popular bit-plane coding paradigm and gives excellent results for lossless compression. Tarp filtering is better for predicting images with wavelet coefficients. Performance of wavelet coders are evaluated in terms of peak signal noise ratio and bit rate for objective quality assessment of reconstructed image. Experiments on test images identified the optimal wavelet encoder combination. The test results show that Cohen-Daubechies-Feaveau 9/7 along with SPIHT encoder yields comparable compression efficiency over other methods.

Keywords: Set Partitioning in Hierarchical Trees(SPIHT), Set Partitioned Embedded bloCK Coder(SPECK), Binary Set Splitting with K-d trees(BISK), Image Compression, Wavelet Transform.

1 Introduction

When considering the field of still image compression, there has been a continuing interest in wavelet based embedded image coders because of their exquisite features, namely high quality at large compression ratios, very fast decoding, progressive transmission etc. Particularly, set-partitioning schemes always have an advantage because of comparatively low-complexity and high performance when compared to other coders, such as those employing vector quantization, Huffman coding etc. A number of hierarchical coding techniques have emerged and all these coding techniques are based on the idea of partitioning the image into sets, and exploiting the hierarchical subband pyramidal structure of the transformed image [1]. Data compression tries to reduce the size of the image by removing the spatial and structural redundancies. This eventually reduces the number of bits required to represent an image [2].

If compression results in exact reproduction of the original image it is called lossless and lossy when we need to incorporate certain quantization [3]. The efficient scheme, Set Partitioning in Hierarchical Trees (SPIHT) using bit-plane paradigm was introduced [1]. A new compression method combining wavelet transform along with SPIHT for 2-D arranged ECG signals was developed and it found better compression with the existing methods [5]. An exhaustive comparative analysis of different compression techniques like SPIHT, SPECK, JPEG2K, EBCOT etc, the recent trends and their applications in the emerging fields of medical science was done [6].Transforming the

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 260–269, 2013.
© Springer International Publishing Switzerland 2013

electromyogram signal by applying discrete cosine transform, wavelet transform on the coefficients and further coding the coefficients using SPIHT encoding resulted in good compromise with percentage root mean square difference, compression ratios etc [7].

A Novel Embedded Set Partitioning and Zero Block Coding for medical image compression showed comparable results over SPECK and JPEG 2000 [8]. A novel compression scheme employing set partitioning in hierarchical trees algorithm (SPIHT), sub-band energy compression (SEC) method on two-dimensional electrocardiogram (2D-ECG) gave comparative results [9]. An efficient compression scheme using Multiwavelet transform with Set Partitioned Embedded bloCK coder algorithm (SPECK) resulted in better compression [10]. Fowler introduced binary set splitting with k-d trees (BISK), an embedded wavelet-based image coder based on the popular bitplane-coding paradigm, BISK is designed specifically for the coding of image objects with arbitrary shape. While other similar algorithms employ quadtree-based set partitioning to code significance-map information, BISK uses a simpler, and more flexible, binary decomposition via k-d trees. BISK can be considered to be a variant of SPECK [11].

A 2-D Tarp filter which estimated the significance probability through an IIR filtering technique was introduced. The estimated significance probability is used to drive a nonadaptive arithmetic coder to compress the information. The Tarp-filter-based image coder can achieve performance comparable to JPEG2000 [18]. SPIHT is an improved version of embedded zero tree wavelet encoding scheme. It has less computational complexity and has non explicit rate distortion optimization [22]. A kind of wavelet transform and bilinear interpolation for image matching was presented, combined with the SPIHT (Set Partitioning In Hierarchical Tree) coding scheme. Experiments were done in comparision with the original SPIHT algorithm, and showed that the new scheme had more prominent advantages in the peak signal to noise ratio, mean-square deviation and histogram [23].

This paper presents the comparison of various wavelet encoders using Cohen-Daubechies-Feauveau-5/3 and Cohen-Daubechies-Feauveau-9/7 wavelet transform. Various test images are tested by first applying wavelet transform and then using the wavelet encoders like SPIHT, SPECK, BISK and TARP filter. Performance of the wavelet coders are evaluated in terms of peak-signal-to-noise ratio (PSNR) and bit rate for objective quality assessment of reconstructed image. Experiments on test images identified the optimal wavelet encoder combination. In this paper, Section 2 provides an overview of wavelet transform and encoding schemes. Section 3 shows the experimental results and Section 4 gives the conclusions arrived with the help of the test samples.

2 Wavelet and Encoding Schemes: Overview

2.1 Wavelet Transform

Discrete wavelet transform (DWT) is a multi-resolutions/multi-frequency representation. The 2D DWT is built with separable orthogonal mother wavelets, having a given regularity. For every iteration of the DWT, the lines of the input image are low-pass filtered with a filter having the impulse response g and high-pass filtered with the filter h. Then the lines of the two images obtained at the output of the two filters are decimated with a factor of 2. Next, the columns of the two images obtained are low-pass filtered

with g and high-pass filtered with h. The columns of those four images are also decimated with a factor of 2. Four new sub-images are generated. The first one, obtained after two low-passes filtering, is named approximation sub-image [17]. The other three types of detail coefficients and basis functions are named: horizontal detail, vertical detail and diagonal detail. The approximation image represents the input for the next iteration. Wavelet transform with 3 levels of decomposition is shown in Fig 1 [21].

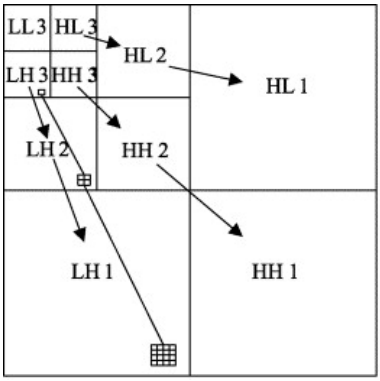


Fig. 1. 2-D DWT with 3 levels of Decomposition

In most cases Wavelet Transform produces floating point coefficients and although this allows perfect reconstruction of the original image. In theory, the use of floating-point arithmetic results in lossy compression. The integer version of the wavelet transform can be obtained using a lifting scheme. Lifting allows wavelet transform to be computed very quickly through a series of individual lifting steps with less memory usage [12].To obtain a 1-D wavelet lifting the input signal is split into odd and even samples followed by a series of predict (P) and update (U) lifting steps to obtain the low pass and high pass coefficients. The LP coefficients are the coarse values and the HP coefficients indicate the detail values in the data. The predict and update steps are also known as dual and primal lifting steps. The inverse transform is obtained by reversing the steps of the forward transform and changing the signs. The forward and inverse lifting steps are shown in Fig 2 and Fig 3 [14]. In the predict step, the odd elements are

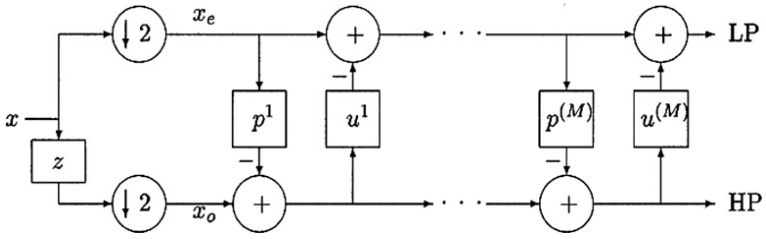


Fig. 2. Forward Lifting step

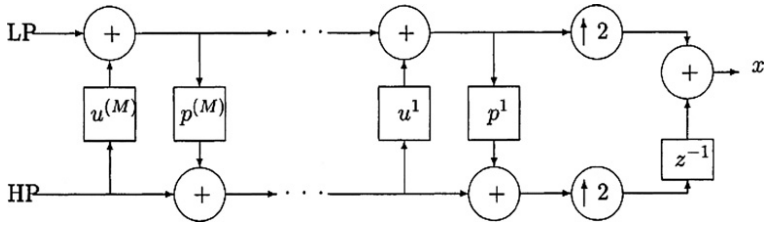


Fig. 3. Inverse Lifting Step

predicted using the even elements. Then the detail coefficient is computed by finding the difference between original odd value and its predicted value. Further the detail coefficient replaces the original odd value and so on. If d represents detail coefficient and s represents the coarse coefficient then, the prediction step can be defined as $d_{n-1} = d_n - P(s_n)$, where P is the prediction operator. Similar to the predict step, the update step can be defined as $s_{n-1} = s_n + U(d_{n-1})$, where U is the update operator [12].

Cohen-Daubechies-Feauveau wavelet are the historically first family of biorthogonal wavelets, which was made popular by Ingrid Daubechies. These are not the same as the orthogonal Daubechies wavelets, and also not very similar in shape and properties. However their construction idea is the same. Cohen-Daubechies-Feauveau 9/7 wavelet transform is commonly regarded to have good performance in image compression and used as a default wavelet transform for JPEG2000 encoding scheme [24]. The predict and update steps required for the wavelets considered in this paper are given in Table 1.

Table 1. Forward transforms for lifting based wavelet transforms

CDF 9/7	$d_1[n]=d_0[n]-203/128\ s_0[n+1]-s_0[n]$
	$s_1[n]=s_0[n]-217/4096\ d_1[n]-d_1[n-1]$
	$d_2[n]=d_1[n]+113/128\ s_1[n+1]+s_1[n]$
	$s_2[n]=s_1[n]+1817/4096\ d_1[n]+d_1[n-1]$
CDF 5/3	$d[n]=d_0[n]-1/2\ s_0[n+1]+s_0[n]$
	$s[n]=s_0[n]+1/4d[n]+d[n-1]$

2.2 SPIHT Encoding

The SPIHT encoding technique was developed by Said and Pearlman in 1996. It uses wavelet sub band decomposition and imposes a quad tree structure across the sub bands to order the transform coefficients. By sending the most important information of the

first ordered coefficients, the information required to reconstruct the image is extremely compact. SPIHT algorithm uses a data structure called the spatial orientation trees (SOT). This particular structure is useful for identifying different scales the correlation between the wavelet coefficients, and also gives full consideration to the correlation of the same scale wavelet coefficients. The algorithm searches each tree, and divides the tree into one of three lists: 1) the list of significant pixels (LSP) containing the coordinates of pixels found to be significant for the current threshold; 2) the list of insignificant pixels (LIP), with pixels that are not significant for the current threshold and 3) the list of insignificant sets (LIS), which contain information about trees that have all the constituent entries to be insignificant at the current threshold. The SPIHT algorithm involves three stages: initialization, sorting pass and refinement pass. At the initialization stage the SPIHT first initializes the threshold according to the maximum value in the wavelet coefficients pyramid, then sets the LSP as an empty list and puts the coordinates of all coefficients in the coarsest level of the wavelet pyramid (LL band) in the LIP and those which have descendants to the LIS. In the sorting pass, the elements in the LIP, followed by the elements in LIS are sorted. For each pixel in the LIP it performs a significance test against the current threshold and outputs the test result (0 or 1) to the resultant bit stream. If a coefficient is significant, its sign is coded and then its coordinate is moved to the LSP. During the sorting pass of LIS, the SPIHT does the significance test for each set in the LIS and outputs the significance information (0 or 1). If a set is significant, it is partitioned into its offspring and leaves. The current threshold is divided by 2 and the sorting and refinement stages are again done until we achieve the target bit-rate [1].

2.3 SPECK Encoding

Consider the SPECK algorithm, in which the quadtree is formed by successive recursive splitting of a subband block (parent) into four quadrants (children). The encoding scheme consists of a sorting pass and a refinement pass. In the sorting pass, when data set S is found significant, it is partitioned into four small child sets, O(S);each of these four child sets is further tested for significance and partitioned until all the significant coefficients are found. In the refinement pass, the coefficients which were found significant in the sorting pass are passed to decoder according to the bit-plane transmission mechanism. The idea behind this is to utilize the clustering of energy found in transformed images and concentrate on those areas of the image which have high energy [4].

2.4 BISK Encoding

k-d Trees. Like quadtrees, k-d trees are a recursive spatial-partitioning data structure. When considering quadtrees, which subdivides a block into four equally sized subblocks, k-d trees insist a binary partitioning; that is, blocks will be divided in two. Several approaches for selecting the location and orientation of the split are available. This approach involves a splitting method in which blocks are divided into approximately equally sized halves, and splitting is done horizontally and vertically alternatively. For example, if the main block is halved vertically, then its two subblocks will be halved horizontally. In particular, if a set of coefficients, S, have decomposition level l(S) in the

k-d tree. Suppose the set S is split, then the two resulting subsets, S1 and S2, will reside at level l(S) + 1 of the k-d tree. The splitting of the set S is done either horizontally or vertically depending on whether l(S) is even or odd, respectively. If noted carefully one can see that a k-d tree can achieve a partition identical to that of a quadtree but will require twice as many levels of subdivision to do so.

The BISK Algorithm. The BISK algorithm initiates by splitting the set of transform coefficients, X, into individual subbands S which are then placed in a list of insignificant sets (LIS). This follows the common bitplane-coding paradigm consisting of sorting and refinement passes. In the sorting pass it determines the significance of a set by comparing the largest opaque-coefficient magnitude contained in the set to the current threshold. When a set has no significant coefficient, then it is placed in the LIS, and, during the sorting pass, a significance test is done for each set in the LIS against the current threshold. If the set becomes significant, then the set is split into two according to the k-d tree decomposition structure. The two new sets are placed into an LIS, recursively tested for significance, and split again if needed. At any time, if a set contains no opaque coefficients, then it is removed from its LIS and discarded [11].

2.5 Tarp Filter Encoding

Tarp filtering consists of coding a binary-valued field by applying a first-order recursive filter to estimate the probability of the next symbol being a 1, and then driving a non-adaptive arithmetic coder with the estimated probability. Tarp filtering scheme employs three 1D filters where each filter runs in a unique direction, the first filter runs from left to right, the second runs from right to left (after processing a full row), and the third runs from top to bottom for each column. The 1D filters, as a result, efficiently implement a 2D convolution, which, in turn, embodies a 2D Parzen windows probability estimate; a parameter, called the learning rate, controls the spread of the Parzen window. Each bitplane of the quantized wavelet coefficients is encoded seperately using the tarp-filter coder, thus making the bitstream non-embedded. An embedded tarp coder was implemented in QccPack [15] by adopting successive-approximation bitplane coding, using tarp filtering to generate probability estimates for the coding of the coefficient significance states [16][17].

3 Experimental Results

Images are compressed by first applying wavelet transform followed by encoding schemes like SPIHT, SPECK, BISK and tarp filter. The parameters that are used for the performance comparison are peak signal-to-noise ratio (PSNR) and bit rate (BR) defined as in (1) and (3) respectively [20].

$$PSNR = 20\log_{10}\{\frac{Maximum\ pixel\ value}{\sqrt{MSE}}\} \tag{1}$$

$$MSE = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}||f(i,j) - g(i,j)||^2 \tag{2}$$

where m,n are the dimensions of the image, g(i, j) denotes the pixel value in the reconstructed image and f(i, j) is the pixel value in the original image. A lower value for MSE means lesser error, and as seen from the inverse relation between the MSE and PSNR, this translates to a high value of PSNR. Logically, a higher value of PSNR is good because it means that the ratio of SNR is higher.

$$BR = \frac{Compressed\ image\ size\ in\ bits}{Total\ number\ of\ pixels} \tag{3}$$

The test results of the encoding schemes are available in Table 2 and 3.

Table 2 shows the test results using Cohen-Daubechies-Feauveau -5/3 wavelet transform and Table 3 show the test results using Cohen-Daubechies-Feauveau-9/7 wavelet

Table 2. Compression results of PSNR in decibels (dB) using Cohen-Daubechies-Feauveau -5/3 wavelet transform

Images	**Encoder**	**Bit rate(bpp)**				
		0.1	0.5	1.0	2.0	3.0
Lena	SPIHT	23.29	24.14	24.25	24.32	24.35
	SPECK	23.32	24.14	24.25	24.32	24.35
	BISK	23.31	24.14	24.25	24.32	24.35
	TARP	23.28	24.12	24.24	24.32	24.35
Barbara	SPIHT	20.74	22.83	23.32	23.50	23.54
	SPECK	20.81	22.86	23.33	23.50	23.54
	BISK	20.83	22.87	23.33	23.50	23.54
	TARP	20.73	22.82	23.29	23.49	23.54
Goldhill	SPIHT	23.69	25.31	25.80	26.11	26.20
	SPECK	23.73	25.31	25.80	26.10	26.20
	BISK	23.73	25.31	25.80	26.10	26.20
	TARP	23.67	25.27	25.77	26.09	26.19

Table 3. Compression results of PSNR in decibels (dB) using Cohen-Daubechies-Feauveau -9/7 wavelet transform

Images	**Encoder**	**Bit rate(bpp)**				
		0.1	0.5	1.0	2.0	3.0
Lena	SPIHT	29.77	37.05	40.25	44.84	**49.91**
	SPECK	29.90	37.06	40.21	44.70	**49.71**
	BISK	29.88	37.03	40.19	44.65	**49.65**
	TARP	29.76	36.72	39.84	44.44	**49.41**
Barbara	SPIHT	24.15	31.25	36.39	42.49	47.58
	SPECK	24.28	31.42	36.41	42.43	47.46
	BISK	24.28	31.44	36.38	42.41	47.43
	TARP	23.93	31.05	35.86	42.35	47.09
Goldhill	SPIHT	27.59	32.90	36.31	41.72	47.12
	SPECK	27.68	32.90	36.29	41.56	46.83
	BISK	27.67	32.91	36.29	41.55	46.80
	TARP	27.67	32.91	36.13	41.37	46.04

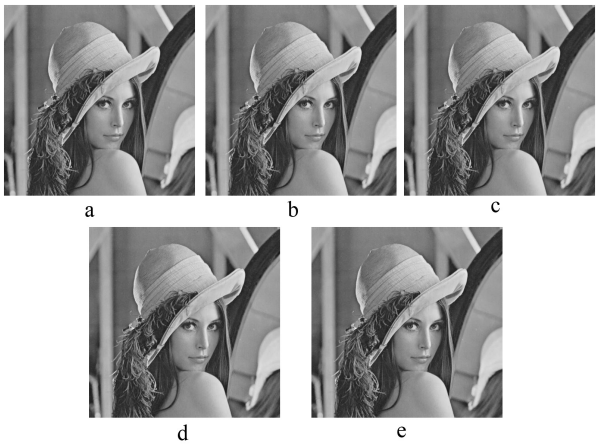


Fig. 4. (a) Original Image, (b) Lena image with SPIHT using CDF-9/7 for a bitrate of 3.0, (c) Lena image with SPECK using CDF-9/7 for a bitrate of 3.0, (d) Lena image with BISK using CDF-9/7 for a bitrate of 3.0, (e) Lena image with TARP using CDF-9/7 for a bitrate of 3.0

Fig. 5. (a) Original image, (b) Barbara image with SPIHT using CDF-9/7 for a bitrate of 3.0, (c) Barbara image with SPECK using CDF-9/7 for a bitrate of 3.0, (d) Barbara image with BISK using CDF-9/7 for a bitrate of 3.0, (e) Barbara image with TARP using CDF-9/7 for a bitrate of 3.0

transform. The number of decomposition levels is set to 4 for better compression. Experiments are done on test images like lena, barbara and goldhill. Various bit rates are given to measure the performance of the encoding schemes like SPIHT, SPECK, BISK and TARP. It can be seen that as the bit rate increases there is an increase in the PSNR value which results in improved quality of the image. Results show that Cohen-Daubechies Feauveau-9/7 along with SPIHT encoding scheme has outperformed well with other methods.

Fig.4. (a) is the original image of Lena used for compression and Fig. 4. (b)-(e) shows the images after applying the different encoding schemes. Fig. 5. (a) is the image of Barbara before applying the compression schemes and Fig. 5. (b)-(e) shows the images after applying the different encoding schemes. Table 3. and Fig.4.(e) shows that a good PSNR value of 49.91 is achieved and quality of image is good when using SPIHT along with Cohen-Daubechies-Feauveau-9/7 for lena image with bitrate 3.0.

4 Conclusion

This paper analyses the performance evaluation of different encoding schemes.Wavelet transforms like Cohen-Daubechies-Feauveau-5/3 and Cohen-Daubechies-Feauveau-9/7 are used with different encoding schemes like SPIHT, SPECK, BISK and TARP. The number of decomposition levels is set to 4. Various test images like Lena, Barbara and GoldHill are used. Experiments on test images are done by computing PSNR and Bit rate. Test results show that Cohen-Daubechies-Feauveau-9/7 wavelet along with SPIHT encoder yields better compression over other methods.

References

1. Pearlman, W., Said, A.: A new fast and efficient Image Codec Based on Set Partitioning in Hierarchical Trees. IEEE Transactions on circuit and systems for video technology 6(3), 243–250 (1996)
2. Rajkumar, P., Mrityunjaya, V.L.: ROI Based Encoding of Medical Images: An Effective Scheme Using Lifting Wavelets and SPIHT for Telemedicine. International Journal of Computation Theory and Engineering 3(3), 338–346 (2011)
3. Sriraam, N., Shyamsunder, R.: 3-D medical image compression using 3-D wavelet coders. Digital Signal Processing 21, 100–109 (2011)
4. Pearlman, W., Islam, A.: Efficient, Low-Complexity Image Coding with a Set-Partitioning Embedded Block Coder, pp. 1–23
5. Goudarzi, M.M., Taheri, A., Pooyan, M.: Efficient Method for ECG Compression Using Two Dimensional Multiwavelet Transform. International Journal of Information and Communication Engineering 2, 8 (2006)
6. Ansari, M.A., Anand, R.S.: Recent trends in image compression and its application in telemedicine and teleconsultation. In: Proceedings of XXXII National Systems Conference, pp. 59–64 (2008)
7. Ntsama, E.P., Pierre, E., Basile, K.I.: Compression Approach of EMG Signal Using 2D Discrete Wavelet and Cosine Transforms. American Journal of Signal Processing 3(1), 10–16 (2013)

8. Rezazadeh, I.M., Moradi, M.H., Nasrabadi, A.M.: Implementing of SPIHT and Sub-band Energy Compression (SEC) Method on Two-Dimensional ECG Compression: A Novel Approach. In: Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference (2005)
9. Geetha, P., Annadurai, S.: Medical image compression using a novel embedded set partitioning significant and zero block coding. The International Arab Journal of Information Technology 5(2), 132–139 (2008)
10. Radhakrishnan, S., Subramaniam, J.: Novel Image Compression Using Multiwavelets with SPECK Algorithm. The International Arab Journal of Information Technology 5, 45–51 (2008)
11. Fowler, J.E.: Shape adaptive coding using binary set splitting with k-d trees. Proceedings of the IEEE International Conference on Image Processing 2, 1301–1304 (2004)
12. Sweldens, W.: The lifting scheme: A construction of second generation wavelets. SIAM J. Math. Anal. 29, 511 (1998)
13. Deever, A.T., Hemami, S.S.: Lossless image compression with projection-based and adaptive reversible integer wavelet transforms. IEEE Trans. Image Process 12, 489–499 (2003)
14. Burrus, C.S., Gopinath, R.A., Guo, H.: Introduction to Wavelets and Wavelet Transforms. Prentice-Hall International (1997)
15. Fowler, J.E.: QccPack: An open-source software library for quantization, compression, and coding. In: Tescher, A.G. (ed.) Applications of Digital Image Processing XXIII, San Diego, CA. Proc. SPIE, vol. 4115, pp. 294–301 (2000)
16. Shah, V.P., Fowler, J.E., Younan, N.H.: Tarp filtering of block-transform coefficients for embedded image coding
17. Tian, C., Hemami, S.S.: An embedded image coding system based on tarp filter with classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 3, pp. 49–52 (2004)
18. Simard, P., Steinkraus, D., Malvar, H.: On-Line Adaptation in Image Coding with a 2-D Tarp Filter, Microsoft Research
19. Lewis, A.S., Knowles, G.: Image compression using 2-d wavelet transform. IEEE Trans. on Image Processing 1(2), 244–250 (1992)
20. Bhaskaran, M., Konstantinides, K.: Image and Video Compression Standards Algorithms and Architectures. Kluwer Academic Publishers (1996)
21. Shih, M., Tseng, D.: A wavelet-based multiresolution edge detection and tracking. Image and Vision Computing (23), 441–451 (2005)
22. Wang, J., Cui, Y.: Coefficient Statistic Based Modified SPIHT Image Compression Algorithm. Advances in Computer Science and Information Engineering (2), 595–600 (2012)
23. Xiao-Hong, Z., Gang, L.: Research of the SPIHT Compression Based on Wavelet Interpolation Matching Image. In: Proceedings of the International Conference, ICAIC 2011, Part II, pp. 1–8 (2011)
24. Abdullah, M.S., Subba Rao, N.: Image Compression using Classical and Lifting based Wavelets. International Journal of Advanced Research in Computer and Communication Engineering 2(8) (2013)

Directional Decomposition for *Odia* Character Recognition

Chandana Mitra[1] and Arun K. Pujari[2]

[1] Sambalpur University Institute of Information Technology (SUIIT), Odisha
chandanamitra@ymail.com
[2] School of Computer & Information Sc., University of Hyderabad, Hyderabad
arunkpujari@uohyd.ac.in

Abstract. Present work aims at analyzing the role of directional features for efficient recognition of printed *Odia* characters. The characteristics of *Odia* scripts that demand for separate rigorous OCR research are identified. Directional features are extracted by directional decomposition of character image and using fixed zoning. The zones are determined based on the input character patterns. Initial experiment with a modest size of 20 features are taken by considering 4 directions and 5 fixed zones yield very promising results. It is shown that these features yield nearly 95% recognition accuracy by multi-class SVM classifiers. High acuracy of recognition justifies the importance of directional decomposition which indirectly captures the stroke information of the script. In another experiment, we consider 164 dimensional feature vectors by taking zones for the entire image. The observation made here can prove to be useful in building an efficient OCR system for *Odia* characters.

Keywords: OCR of Indic Scripts, Feature extraction, Directional decomposition, Odia scripts.

1 Introduction

Automatic character recognition is becoming extremely important in digital preservation of cultural heritage. The importance of OCR for Indian languages cannot be underestimated as massive volumes of print-media resources representing the cultural and historical heritage of India are available in several Indian scripts. Several recognition techniques for printed characters of different Indian scripts have been reported in literature [3, 7-11]. Like Devanagari and Bangla scripts, *Odia* script derives its origin from Brahmi script. Though there have been substantial research in recognition of Indic Characters, there has not been much of attentions paid to *Odia* characters except few isolated study [3] or generalized investigation [7]. One of the main reasons for lack of rigorous study is the assumption that a field-proven technology for any other Indian languages can perhaps be suitably adapted to recognize Odia scripts. This assumption is not valid and we justify this to some extent as we describe our results here.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 270–278, 2013.
© Springer International Publishing Switzerland 2013

The choice of feature set and extraction of features play very important roles in designing successful OCR system. Over the years, innumerable feature extraction methods have been suggested for OCR of different languages. Some of the review articles give accounts of trends in feature extraction of techniques [12]. Directional features are shown to be very useful for recognition of Chinese and Japanese handwritten characters [1, 3-6]. There has been detailed analysis of effectiveness of directional features for Kanji characters [4]. Many algorithms have been developed to decompose the character-images into directional subimages incorporating the applicable criteria that follow. Many algorithms have been developed which decompose the character-images into directional subimages.

In this paper, our objective is to demonstrate that a simple but novel directional decomposition technique is effective for recognition of printed *Odia* characters. *Odia* characters are though circular in nature most of the distinguishing information occurs in non-circular portions. The relative position and orientation of linear strokes can be cleverly used to distinguish individual characters. The objective of the present work is to demonstrate the efficacy of elegant directional features for successful classification of *Odia* printed character. In the proposed approach, the preprocessing steps of skeletonizing and contour extraction are avoided.

In section 2, we discuss the specific characteristics of *Odia* script that demand a separate OCR technology. We introduce the concept of directional decomposition of character images. Several characteristics of directional decomposition and its relation to Mathematical Morphology are also discussed. We also show the relevance of directional decomposition to *Odia* characters.

2 ODIA Scripts

Odia (till recently, referred to as Oriya) script, derived from Brahmi script, is predominantly used in Odisha state. The *Odia* script is circular in nature and there is no accompanying line at the top of a character. It is believed that the circular shape evolved from the need to write on palm leaves with a pointed stylus, which has a tendency to tear for horizontal or vertical strokes. With the advent of modern printing technology, the style of printed scripts has undergone a seachange. In the present study, we consider present style of printing of *Odia* text. The following is the set of vowels of *Odia* scripts.

ଅ ଆ ଇ ଈ ଉ ଊ ଏ ଐ ଓ ଔ

The set of consonants of *Odia* script is the following.

କ ଖ ଗ ଘ ଙ ଚ ଛ ଜ ଝ ଞ ଟ ଠ ଡ଼ ଢ଼ ଣ ତ ଥ ଦ ଧ ନ

ପ ଫ ବ ଭ ମ ଯ ୟ ର ଲ ଳ ସ ଶ ଷ ହ କ୍ଷ ଜ୍ଞ

Vowels may appear in pure form as an independent vowel in any part of a word or these can be attached as modifiers to the top, bottom, left or right-side of consonants. In ancient style of writing, sometimes a consonant and its attached vowel is a different

atomic symbol and is not simply a concatenation of a vowel symbol around the consonant. Different vowel symbols attached to ka in modern style of printing are listed below.

କା, କି, କୀ, କୁ, କୂ, କେ, କୈ, କୋ, କୌ, କଃ, କଁ, କୃ

In addition to consonants and vowels, there are very large number of atomic symbols representing different consonant-vowel and consonant-consonant combinations. *Akshara* is considered to be the atomic symbol in Indic scripts and an *Akshara* can be an independent vowel, a consonant, a consonant with vowel modifier or a combination of multiple consonants and vowels. In *Odia*, when an *aksharik* symbol represents the combination of multiple consonants, it is not possible to separate the constituent consonants. In some cases two consonants are combined by writing one below the other such that the secondary consonant is of smaller size (e.g., Sh + T, ଷ + ଟ as ଷ୍ଟ). In some cases a part of the secondary consonant is attached to the bottom of the primary consonant (e.g., L +k, ଳ + କ as ଳ୍କ) or, altogether a new symbol is attached to the primary consonant. The same symbol may represent different secondary consonants in different contexts (e.g., N+D and n+t, ଣ୍ଡ, ନ୍ତ). In another case of combination of two consonants, the resulting akshara does not retain the shape of any of the constituent consonants. In some cases the primary consonant is reduced in size to facilitate uniform size for isolated consonant and combined consonants. In other cases, the size of the primary consonant is not altered. For example, in cases of m+t ମ୍ତ, and m+bh ମ୍ଭ, having same primary consonant m ମ in one font but are of different sizes. There are 415 *aksharas* in current printing style of *Odia* and we categorize consonant-consonant *aksharas* as follows.

- *Aksharas* which are a combination of two consonants but not preserving the shape of any of the constituent consonants.

  ଙ୍କ ଙ୍ଖ ଙ୍ଗ ଙ୍ଘ ଞ୍ଚ ଞ୍ଛ ଞ୍ଜ ଞ୍ଝ ଣ୍ଟ ଣ୍ଠ ଣ୍ଡ ଣ୍ଢ ନ୍ତ ନ୍ଥ ନ୍ଦ ନ୍ଧ ମ୍ପ ମ୍ଫ ମ୍ବ ମ୍ଭ

- *Aksharas* having one consonant modified with a symbol which does not resemble the second consonant. While applying the modifier, the size of the first consonant is reduced in scale. The modifiers appear normally at the right-bottom part of the first consonant. Most often, a modifying symbol can be unambiguously mapped to a single (or more) second consonant(s). However, there are few symbols which are used for two different consonants while used in different contexts.

  ଶ୍ଚ ଶ୍ଛ ନ୍ଚ ନ୍ଛ ମ୍ଚ ମ୍ଛ କ୍ର ଖ୍ର ଗ୍ର ଘ୍ର ପ୍ର ଫ୍ର ବ୍ର ଭ୍ର ମ୍ର ଶ୍ର

- *Aksharas* with a primary consonant that is modified by a modifier resembling a portion of second consonant. One can see that mostly the lower portion of second consonant is retained. Thus the semi-circular curve at the top, which is common in most of the consonants, is dropped while combining two consonants. The size of the first consonant is scaled down.

  ଜ୍ଞ ବ୍ଳ ଳ୍କ ଳ୍ଗ ଳ୍ଜ ସ୍କ ସ୍ଖ ସ୍ଟ ସ୍ଠ ସ୍ପ ସ୍ଫ ହ୍ମ ହ୍ୟ ହ୍ର ହ୍ଲ ହ୍ୱ

- *Aksharas* preserving the shape of both the consonant written one below the other. The consonant written below is smaller in size than that of the top consonant. Both the consonants are smaller than the respective unmodified sizes in the given font. The second consonant is attached to the bottom-right portion of the first consonant.

ଦ୍ନ ଦ୍ମ ଦ୍ଲ ଦ୍ବ ଦ୍ଭ ଦ୍ୱ ଦ୍ଧ ଦ୍ଦ ଦ୍ଥ ଦ୍ତ ଦ୍ଣ ଦ୍ଡ ଦ୍ଠ ଦ୍ଟ ଦ୍ର

- The size of the first consonant is preserved and modified with a symbol that does not resemble any consonant. The modifier is attached to bottom, mostly bottom-right of the first consonant.

ମ୍ଜ ମ୍ପ ମ୍ମ ମ୍ଗ ମ୍ଫ ଧ୍ନ ନ୍ଧ ଶ୍ଚ ତ୍ତ ହ୍ମ

It is clear from the foregoing discussion that routine application of OCR systems for other Devnagari-derived or Dravidian-derived scripts will not be useful for recognition of *Odia* characters. The distinguishing features ought to be extracted keeping in mind the characteristics of *Odia Aksharas*. It is evident from foregoing discussion that any *Odia* OCR must be able to capture features which are scale invariant, zone-based, and not sensitive to semi-circular shape at the top.

3 Directional Decomposition

The present study attempts to examine the effectiveness of directional decomposition for recognition of Odia script. There are several definitions of directional features. In the present scheme, it is to partition the set of foreground pixels of the character image into k images, where k is the number of directions. The partition is done based on the occurrences of directional features and the process is called directional decomposition. These images are used to extract numeric features. For convenience, we limit our study to four directions- vertical, horizontal, right (southwest to northeast) and left (southeast to northwest). Consider a foreground pixel $p(i,j)$. The directional stretch at $p(i,j)$ in direction d is defined as the number of contiguous foreground pixels in direction d through p and is denoted as $stretch(p(i,j), d)$. Figure 1 illustrates the stretch of a point in four directions.

Fig. 1. Illustration of stretch of a pixel in four directions

A pixel $p(i,j)$is put into bin $d^* = \text{argmax}_d(stretch\ (p(i,j), d))$. There are four bins corresponding four directions and every pixel is in exactly one bin. In a sense, the directional decomposition of a binary image is to decompose the mage into four images such that each of the decomposed images has set of foreground pixels having the

longest stretch in the respective direction. In Figure 1, the pixel under study falls to the bin of left-diagonal bin. Figure 2 shows the results of directional decompositions of character image of the alphabet *e*.

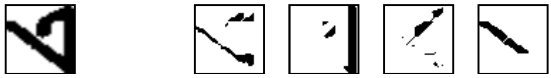


Fig. 2. Directional decomposition in 4 directions a. original image, b. horizontal component, c. vertical, d. right and e. left

One of the first attempts to analyze the performance of directional features is reported in [4]. Projecting a directional vector to relevant directional axis is a popular method. Very recently, there have been attempts to decompose the whole image into directional subpatterns to identify predominant strokes. In [6], four different strategies of cellular directional decomposition are proposed. In [5], directional decomposition with elastic mesh is used for recognition of handwritten Chinese characters. The method discussed in the present work is derived from stroke-base directional decomposition technique proposed in [5]. There are many interesting properties of directional decomposition. We analyze some of these properties to provide insight into such decompositions and its uses for efficient recognition of characters.

The directional decomposition can be related to mathematical morphology and can be generalized. The main idea is to determine the maximal scale that a structuring element centered at a given foreground pixel is inside the input image. The structuring element S is taken with different orientations, say k. in this context, by $stretch(S(i,j), d) = \lambda$ we mean that λ is largest scaling factor so that the structuring element S with orientation d and centered at pixel $p(i,j)$ so that it is completely inside the original image. The input image is decomposed into k subimages such that each pixel becomes of the part of d^{th} subimage if the d^{th} orientation of the structuring element centered at the pixel yields maximal scale. That is, pixel $p(i,j)$ is put in bin d* if $d^* = \text{argmax}_d(stretch\ (p(i,j), d))$. In the present study, we consider the following structuring element with 4 orientations, 0^0, 45^0, 90^0, and 135^0.

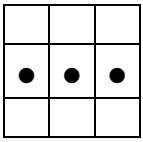


Fig. 3. Structuring element with 0^0 orientation

By having different the structuring elements one can decompose the image to analyze several geometric properties such as convexity, connectedness etc. [13]. In the present study, we have experimented with several different structuring elements and shall report in detail in a forthcoming paper. Upper and lower semi-circular curves

can be uniquely captured by horizontal components of directional decomposition. Figures 4 a-c give the horizontal components of 3 *Odia* characters, *Tha*, *ba*, and *A*. One can see that upper and lower parts of *Tha* and *ba* are sets of horizontal lines. Similarly, the concavity of A in the upper part is also captured.

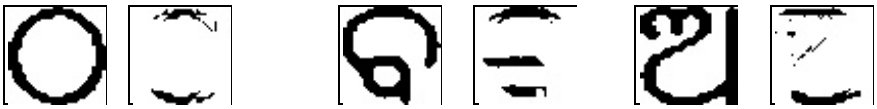


Fig. 4. a. Horizontal component of *Tha* b. *ba* c. *A* and respective input images

Fig. 5. Vertical features a. for *A* and b. for *Ja*

The vertical primitives of a character are separated out by the vertical component during directional decomposition. To see this, consider characters A and Ja (Figure 5a-b). In Figures 6 a-b, it can e seen that left slant and right-slant strokes are identified for SHa and Ya, respectively.

Fig. 6. a. left slant of *Sha* is extracted by left component b. right slant of *Ja* is captured by right component.

Zoning

We propose to extract features from the components resulting from directional decomposition in two different ways. In the first method, we subdivide the image of size N×N to equal blocks of size 8×8. For each such block, the density of foreground pixel is calculated in each of *d* component images. In our experiments, the images are normalized to 48×48 and hence we get 36 such blocks. For directional decomposition with 4 directions, i.e. *d=4*, we have 144 features. It may be noted that these 36 blocks are independent of the actual image patterns. Thus images of same character in two different fonts may yield different set of features.

In order to overcome the sensitivity of fonts, the second set of zoning is taken based on actual span of foreground pixels in input images. Five zones (Figure 7) in the character images are identified. These zones are chosen based on the discussion in Section 2. For instance, bottom-right corner of the character is the region where the second consonant or modifiers are located. Let x_1 (x_2) be the topmost (bottom-most, respectively) row of foreground pixels of the character image. Thus, x_2-x_1 is the height of the character image. Similarly, let y_1 and y_2 be the leftmost and rightmost columns, respectively, of the character image. Zone 1 is the left-bottom portion of the image and characters looks

similar except for the difference in shape in this region such as *ka*, *ba*, and *da*. Zone 2 attempts to extract the density at top-right corner of the foreground part of the image. It is limited to top half rows and 3rd quarter of columns. In addition to 3rd quarter we also take 3 additional columns to take care of vertical lines. Zone 3 is limited to top 1/3rd of rows, first 3 quarters of columns. Zone 4 consists of a rectangular region of middle part of the character image. Zone 5 is meant to capture the modifiers and consonant symbols appearing in the lower-right portion of the image. It is the last 33% of rows and 40% of columns. Table 1 gives the detailed definitions of these zones.

Table 1. Description of zones

	Start row	Las row	Start column	Last column
Zone 1	$\frac{(x_2 - x_1)}{2}$	x_2	y_1	$y_2 - \frac{(y_2 - y_1)}{4}$
Zone 2	x_1	$\frac{(x_2 + x_1)}{2}$	$\frac{(y_2 + y_1)}{2}$	$y_2 - \frac{(y_2 - y_1)}{4} + 3$
Zone 3	x_1	$x_1 + \frac{x_2 - x_1}{3}$	y_1	$y_2 - \frac{(y_2 - y_1)}{4}$
Zone 4	$x_2 + \frac{x_2 - x_1}{4}$	$\frac{(x_2 + x_1)}{2}$	y_1	$y_2 - \frac{y_2 - y_1}{4}$
Zone 5	$x_2 - \frac{x_2 - x_1}{3}$	x_2	$y_1 + 0.4\frac{y_2 - y_1}{2} + 4$	y_2

Fig. 7. Fixed Mesh- Five regions as outlined

We get a 20-dimensional feature vector with foreground pixel densities in 5 zones for 4 components. Zoning for the purpose of feature extraction is also known as *meshing*. The image area is subdivided into blocks where the directions features are used to extract some numeric quantities. There are two kinds of meshing methods- *fixed* meshing and *elastic* meshing [5]. In fixed-meshing method, the position of meshing is fixed for all character images. In Elastic meshing method, a set of position-adaptable meshes are designed according to a specific instance of the image. The proposed zoning can be viewed as fixed meshing method. Elastic meshing is more appropriate for handwritten OCR.

4 Experimental Results

Experiments were carried out to study whether the feature set extracted from directional decomposition can be useful for classification of *Odia* characters. A set of printed pages obtained from different sources (mostly from literary periodicals) are

scanned at a resolution of 300dpi. Connected component method is used to segment individual symbols. The vowel symbols attached to any alphabet are separated out. Most of these attached vowel symbols are not connected to the associated consonant and hence are separated out easily by connected component algorithm. We do not consider these vowel symbols in the present study. Each image is normalized to 48 ×48 size binary image.

We use C++ library LibSVM (version 2.35) from Chang and Lin [2] as the classifier. We use multi-class classifier with RBF kernel. The hyper-parameters were tuned by experimenting with different values of C and γ by a grid search over the two-dimensional parameter space (C; γ)— C ranges from 2^{-5} to 2^{12} and γ from 2^{-10} to 2^{5}.

In first set of experiments, for each of image, a 20-dimensional feature vector is computed by directional decomposition and fixed meshing. Our dataset consists of nearly 2936 images in different fonts of 71 *aksharas* with unequal frequencies. The highest frequency is 183 (for *ra*) and the least frequency is 11 for *kta*. Aksharas with very low frequencies are ignored. The training set consists of 2000 images and we have two separate test sets. The first test set is the remaining 936 symbols drawn from the same documents from which the training set is generated. The second set of test set is from an unknown source printed in different font.

Experiments were carried out for different combination values of *g* ranging from 10-way to 20-way cross validation. We observe that when C= 32, γ = 0.5, a 16-way cross-validation yields accuracy of 98.2511%. For these parameters, the accuracy for test data set was 95.5665% when the test data is a scanned page of the similar source as the training data. For the scanned data of a page which is not part of the training data the accuracy of the classification was 80.1255.%. When C= 0.5 and γ = 0.5, the accuracy for the second test data increased to 86.5063%. For this setting of hyperparameters, the accuracy for first test set was 94.0887%. During our experiments, it is realized that there are certain pairs of *aksharas* which are provably indistinguishable by directional features. For example, two symbols nka and nga(ଙ୍କ ଙ୍ଗ) have almost identical directional feature vectors. There are approximately 9 pairs of symbols (as given below) falling into this category. In another round of experiments, we combined each of these pairs as single symbol in the training set. With new training set, multiclass classification by LibSVM gives 96.15% accuracy for the second set of test data.

In the second set of experiments, we took 164 features, 144 features from 36 8×8 blocks and 20 features from 5 fixed meshing. With 164-dimensional feature vectors, we trained multi-class SVM with different set of training sets and tested with different sets of test data. The test data is drawn from a different set of documents. Table 2 summarizes the results.

Table 2. Accuracy of recognition for different sizes of training set and test sets

Size of the training set	Test set 1 Size 350	Test set 2 Size 400	Test set 3 Size 500	Test set 4 Size 700	Test set 5 Size 956
1500	24	48.75	99	88.64	71.65
1800	64.28	74	99	99.37	86.30
2145	**98.57**	**98.5**	**99**	**99.37**	**98.85**

From these experimental analyses, we conclude that directional decomposition is a useful feature for capturing salient characteristics of *odia* script and recognition can be very accurate.

5 Conclusions

In the present work, it is shown that features extracted by directional decomposition are useful in recognition of printed *Odia* characters. This is due to the fact that the feature-set is able to capture the stroke characteristics of Odia scripts. It is shown that a modest size of features (20 in the present case) and straight-forward use of SVM classifier can lead to high recognition accuracy. The outcome of present study can, in no way, claim to be a full-fledged recognizer, it certainly helps to move a step toward building an efficient OCR system for printed Odia script. It is also discussed here that the inherent properties of *Odia* script demands for rigorous independent OCR research.

References

[1] Bai, Z.-L., Huo, Q.: A study on the use of 8- Directional features for online handwritten character recognition. In: ICDAR 2005, pp. 262–266 (2005)
[2] Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), Software available at: `http://www.csie.ntu.edu.tw/cjlin/libsvmS`
[3] Chaudhuri, B.B., Pal, U., Mitra, M.: Automatic recognition of printed Oriya script. In: SADHANA, Part 1, vol. 27, pp. 23–34 (2002)
[4] Fujisawa, H., Liu, C.-L.: Directional Pattern Matching for Character Recognition Revisited. In: ICDAR, pp. 794–798. IEEE Computer Society (2003)
[5] Gao, X., Jin, L.-W., Yin, J.-X., Huang, J.-C.: A New Stroke-Based Directional Feature Extraction Approach for Handwritten Chinese Character Recognition. In: ICDAR 2010, p. 635 (2001)
[6] Harit, G., Chaudhury, S., Garg, R.: GFG-Based Compression and Retrieval of Document Images in Indian Scripts. In: Govindraju, Setlu (eds.) Guide to OCR for Indic Scripts, Advances in Pattern Recognition, pp. 269–284 (2010)
[7] Jin, L., Wei, G.: Handwritten Chinese character recognition with directional decomposition of cellular features. Jr of Circuit, System and Computers 8, 517–524 (1998)
[8] Negi, A., Bhagvati, C., Krishna, B.: An OCR system for Telugu. In: ICDAR 2001, pp. 1110–1114 (2001)
[9] Pal, U., Chaudhuri, B.B.: Indian script character recognition-A survey. Pattern Recognition 37, 1887–1899 (2004)
[10] Pal, U., Sharma, N., Wakabayashi, T., Kimura, F.: OffLine Handwritten Character Recognition of Devanagari Script. In: Proc. 9th International Conference on Document Analysis and Recognition, pp. 496–500 (2007)
[11] Pujari, A.K., Naidu, C.D., Jinga, B.C.: An adaptive and intelligent character recognizer for Telugu scripts using multiresolution analysis and associative measures. Image Vision Comput. 22(14), 1221–1227 (2002)
[12] Trier, O.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition-A survey. Pattern Recognition 29(4), 641–662 (1996)
[13] Yu, L., Wang, R.: Shape representation based on mathematical morphology. Pattern Recognition Letters 26, 1354–1362 (2005)

Efficient Touching Text Line Segmentation in Tamil Script Using Horizontal Projection

Thangairulappan Kathirvalavakumar* and M. Karthigai Selvi

Research Center in Computer Science
V.H.N.S.N College (Autonomous), Virudhunagar-626 001, India
kathirvalavakumar@yahoo.com, karthigavishaal@gmail.com

Abstract. In this paper an efficient method has been proposed to segment a document of machine printed Tamil sources into text lines. Because of the interfering lines, text line segmentation remain a problem. Standard Horizontal projection method can not segment the lines which are overlapped or touched. But the proposed method uses horizontal projection technique to solve the problem of line overlapping and over segmentation. Experimental results show that 100% accuracy is obtained from the line segmentation process which involves Tamil language document with different sizes and different fonts with line overlapping.

Keywords: Tamil document, Line Segmentation, Projection profile, Overlapping lines.

1 Introduction

People generally store important information by writing on paper for retrieving at a later stage. Paper is still a convenient and feasible way of storing data in the form of handwritten script or printed text. Machine Printed documents have motivated researchers to develop binarization and enhancement algorithms suitable for the challenges such as recognition, skew detection, and page/line segmentation. Various document image segmentation techniques have been proposed in the literature. Pal et al. [15] have proposed a method to separate lines from multi Indian script document with the features like existence of headline, number and position of peaks in horizontal projections and water reservoir. Pal and Chaudhuri [13] have discussed the concept of zoning for line segmentation. Likforman-Sulem et al.[5] have reported several techniques namely Projection profile, Smearing, Grouping, Hough transform, Repulsive-Attractive network and Stochastic for text line segmentation. A conventional technique, global horizontal projection analysis of black pixels has been utilized by Dongre and Mankar [17] for text line segmentation.

Utpal Garain and Chaudhuri[14] have proposed a method for identification and segmentation of touching characters based on fuzzy multifactorial analysis. Karthik et al.[4] have investigated the use of convex optimization methods,

* Corresponding author. The work of T.Kathirvalavakumar is supported by University Grants Commission for Major Research Project, Government of India.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 279–288, 2013.
© Springer International Publishing Switzerland 2013

selective local/global segmentation algorithm and fast global minimization algorithm for simultaneous binarization and segmentation. Dhanya et al.[1] have performed line segmentation by identifying the valley points in the projection profile taken along the rows of the image. Soujanya et al.[9] have proposed different text line segmentation algorithms like Projection Profiles, Run length smearing and Adaptive Run length smearing on low quality documents. Manmatha and Rothfeder [6] have developed State of the art segmentation techniques like gap metrics for highly constrained documents like bank checks and postal addresses. Nikolaos et al.[7] have proposed a method which is a combination of different segmentation techniques to generate improved segmentation results.

Siromony et al.[11] have described a method for recognition of machine printed Tamil characters using an encoded character string dictionary. Jindal et al. [2] have proposed a complete solution for segmenting touching characters in a printed Guruki script. Premaratne and Bigun [8] have proposed a method to recognize Sinhala script characters directly using a standard alphabet as the basis without the need for segmentation into basic components. Vijay Kumar and Sengar [16] have described line and word segmentation method for printed text in Gurmukhi script. Jindal et al.[3] have proposed a solution for segmenting horizontally overlapping lines in eight most widely used printed Indian scripts. Sanjeev Kunte and Sudhaker Samuel [10] have proposed two-stage method for segmenting Kannada Characters which combines connected component analysis and projection profile. Sridevi and Subashini [12] have proposed two methods one for line segmentation and another for character segmentation. First method uses projection profile and PSO for line segmentation but it couldn't segment the touching lines. In the second method, combination of connected components along with nearest neighborhood concept is used to segment the characters. But it couldn't segment the characters if the spaces between the basic characters and modifiers are more.

New Line segmentation method has been proposed in this paper for printed Tamil documents using horizontal projection technique which involves simple procedure. It gives solution for line overlapping and over segmentation and separate each line of a document into separate lines eventhough the document involves different fonts and styles. The rest of the paper is organized as follows. section 2 describes the Characteristics of Tamil Script, section 3 describes the proposed method and in section 4, the experimental results are discussed.

2 Characteristics of Tamil Script

Tamil script consists of 12 vowels, 18 consonants and one special character the ayutam. The vowels and consonants are combined to form 216 compound characters, giving a total of 247 characters.Tamil vowels are called uyireluttu. The vowels are classified into kuril and nedil. Some vowels require the basic shape of the consonant to be altered in a way that is specific to that vowel. Others are written by adding a vowel-specific suffix to the consonant, yet others a prefix, and finally some vowels require adding both a prefix and a suffix to

the consonant. Tamil consonants are known as meyyeluttu. The consonants are classified into three categories vallinam, mellinam and itayinam. Modification of a character is carried out by simply adding one or more modifier symbols before/after/above/below the character without affecting the general shape. Some of the modifiers touches the upper Zone. Some of the modifiers touches the lower Zone. Sometimes lower zone modifiers of a line touches the upper zone modifiers of the next line, thus producing multiple horizontally overlapping lines. Figure 1. shows some examples for upper and lower modifiers.

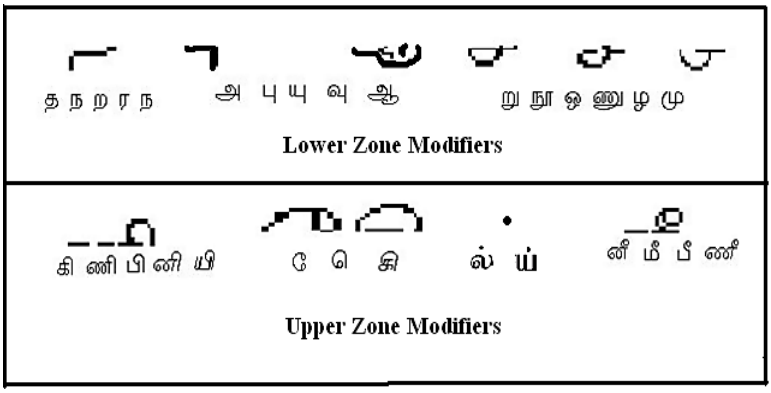


Fig. 1. Modifiers of Lower and Upper zones

Text Line Structure

A text line may be partitioned into three zones. The upper zone denotes the portions above the meanline, the middle zone covers the portion of basic(and compound) characters below the meanline and the lower zone is the portion below the baseline. An imaginary line, where the uppermost points of characters of a text line lie, is referred as meanline. An imaginary line, where the lowermost points of characters of a text line lie, is referred as baseline. Characters which lies above the meanline are ascenders and which lies below the baseline are descenders. Overlapping components of descenders and ascenders located in the region of an adjacent line. Upper line which joins the top of ascenders. Lower line which joins the bottom of descenders. Categories of characters are characters which lies between upper zone and middle zone, characters which lie in the middle zone, characters which lies between middle zone and lower zone and characters which covers all upper, middle and lower zone. Figure 2 shows the text line structure of Tamil script.

3 Proposed Methodology

Segmentation is a process which is used to split the document images into lines. The problem of horizontally overlapping lines makes the line segmentation more difficult. In printed Tamil script, sometimes lower zone characters of a line touches the upper zone characters of next line, thus producing multiple horizontally overlapping lines. It becomes difficult to estimate the exact position

of a row which segments a line from the next line because of horizontally overlapping. A new method has been proposed to segment this kind of document into individual lines.

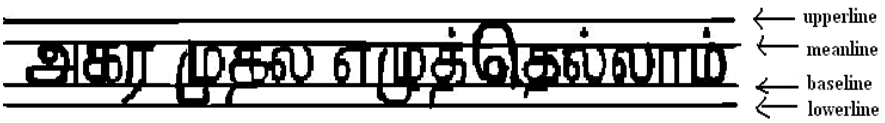


Fig. 2. Text line structure

Binarization is a preprocess which converts a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by selecting a global threshold that separates the foreground from background. Each pixel of a image becomes 1 if it is greater than the threshold and becomes 0 otherwise. Here a threshold of 158 is chosen. Binarization to be performed as a preprocess before the line segmentation process if the image is in a gray scale form. Consider the size of given binary image as M x N where M is the rows namely height and N is the columns namely width of the image. Horizontal projection of the image is defined as :

$$HP(i), i = 1, 2, 3, , M,$$

where HP(i) is the total number of black pixels in i^{th} horizontal row.

Horizontal projection of each row of the image is computed. Collection of consecutive horizontal rows containing at least one pixel are strips. The strips may contain single line of text or multiple lines of text with overlapping. The strips may contain upper zone or lower zone modifier(s). It has been observed that the characteristic of horizontal projection of initial and final rows of a text line are with lesser values than other horizontal projection values of the text line as those rows involve only modifier. The upper and lower zone of a text line are with lesser projection values than the middle zone of the text line. If even number of text lines are overlapped in the strip then mid portion of horizontal projection of the strip are to be lesser than average of all horizontal projection of a strip. If odd number of text lines are overlapped in the strip then each one third portion of horizontal projection of the strip are to be lesser than average of all horizontal projection of a strip. If more than one line is overlapped in the strip then ending of a text line and starting of a next line are with lesser horizontal projection values. Even though lines are overlapped, the horizontal projection of mid point of those two lines is lesser than other horizontal projection of those two lines except the horizontal projection of the rows contain upper and lower modifiers. Finding minimum horizontal projection value in the end portion of a line and starting portion of a followed line give the index which split the overlapped lines into two strips.

Modifiers can be separated from the middle zone of a text whenever horizontal projection value(s) is(are) zero between middle and upper zone of a text and/or middle and lower zone of a text which is called as over segmentation. Over segmentation of upper and lower modifiers of a text need joining operation

for concatenating body of the text with these modifiers. In the strip, instead of overlapping lines the upper or lower zone modifier can be there. It can be identified by finding the number of rows in the strip. Compare to middle zone of text, lower and upper zone modifiers occupy lesser number of rows. A threshold to be used to identify strip with lower modifier or upper modifier.

Algorithm

Step 1: Binarization

• Convert given gray scale image into binary image using standard Otsu's threshold method.

Step 2: Segmentation

• Find the horizontal projection (HP) for each row by counting 'on' state pixels.

$$Find\ \ Threshold(T) = (\frac{Sum\ \ of\ \ non\ \ zero\ \ HP\ \ of\ \ document}{Number\ \ of\ \ strips}) * (\frac{1}{3})$$

•Identify a row with HP=0.

Step 2.1 : Checking for line overlapping

• consider all rows between two consecutive white spaces, where white space means adjacent HP are zero, and treat it as strip S1.
• Find average HP of S1 namely AP.
• If Number of HP in S1 <= T, then this strip is with only modifier(s), it needs joining operation.
• If middle HP of S1 > AP then S1 is not with even number of line overlapping else S1 is with even number of line overlapping.
• If HP of the one third position of the S1 > AP then S1 is not with odd number of line overlapping else S1 is with odd number of line overlapping .
• If even or odd number of lines are not overlapped then S1 is with single line.

Step 2.2 : Procedure for line overlapping

• Find a row with minimum HP, namely MHP, from S1 without considering first and last four HPs.
• Treat from the first row of S1 to the row MHP as strip S2.
• Treat from the row MHP+1 to last HP of S1 as strip S3.
• Apply line overlapping procedure for S2 and S3 until every strip is with a single line.

Step 3: Joining Procedure

• Assume S2 is the strip with modifier, S1 is the strip immediately above S2 and S3 is the strip immediately below S2.

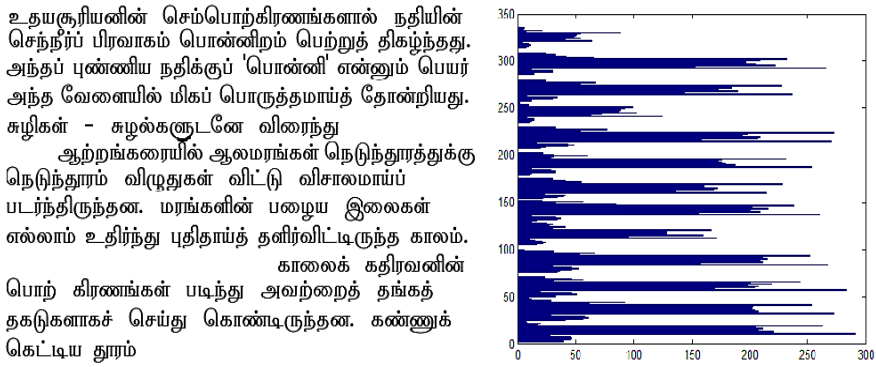
• If number of rows between S1 and S2 is lesser than number of rows between S2 and S3 then join S1 and S2 else join S2 and S3 by removing rows with HP=0.

• If the strip S2 is not preceded by any other strip then S2 to be joined with S3.

• If the strip S2 is not succeeded with any other strip then S2 to be joined with S1.

4 Experimental Result

Different images with different fonts,styles, overlapping are considered for simulating the proposed algorithm for printed Tamil language script. The proposed algorithm also joints together the broken components of an over segmented line. In this manuscript we have drawn histogram to show the text line overlapping. Figure 3a contains components of line need joint operation, strips with odd



(a) (b)

Fig. 3. (a) Original image for Document1 and (b) Histogram of Document1

and even number of overlapping lines. Figure 3b shows the histogram of figure 3a. Figure 3b shows that the image is with overlapping lines and portion of a text need joining operation. The resultant of figure 3a is shown in figure 4a and figure 4b. Figure 4a shows strips with over segmented line, overlapped lines and segmented lines obtained by the standard horizontal projection method. Figure 4b shows the final segmented lines obtained from the procedure. Figure 5a is another example figure contains lines with different styles and fonts and figure 5b shows its corresponding histogram. Strips obtained from figure 5a with standard horizontal projection are shown in figure 6a. Figure 6b shows the resultant lines obtained from the proposed method for the example figure 5a. Figure 7a is an example figure with over segmented lines and 7b is its histogram. Figure 8a shows

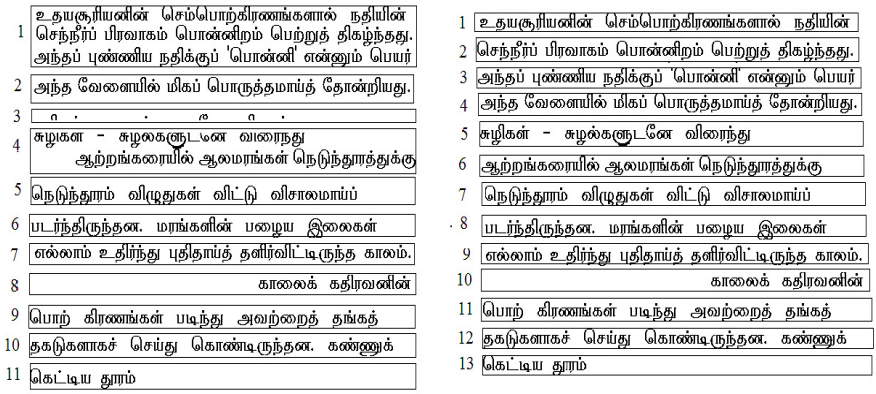


Fig. 4. (a) Various strips in Document 1 using Horizontal projection technique (b) Different lines identified using Proposed Method

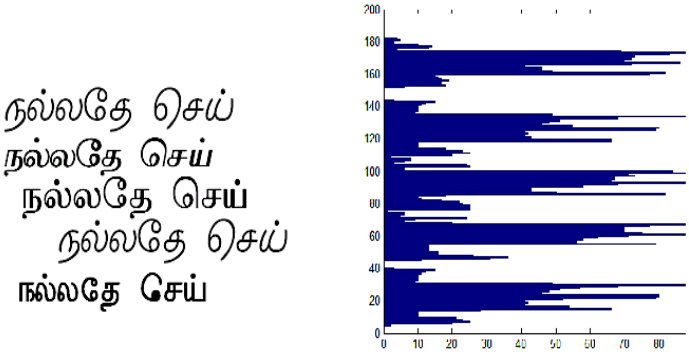


Fig. 5. (a) Original image for Document 2 and (b) Histogram of Document 2

strips obtained by the standard horizontal projection method for the figure 7a and figure 8b shows the final segmentated lines obtained from the figure 7a. The numbers 1,2,3,... in figure-4, figure-6 and figure-8 represent the strip numbers. The results obtained for the considered examples are illustrated in Table 1. The third column of Table 1 shows obtained strips and overlapped lines in each of the strip.

All the results show that the proposed algorithm accurately split all the text lines properly even though the considered images for simulation are with different font, styles, line overlapping and upper zone modifier with over segmentation. In this simulation, to split the overlapped lines, 4 rows from starting of a strip and

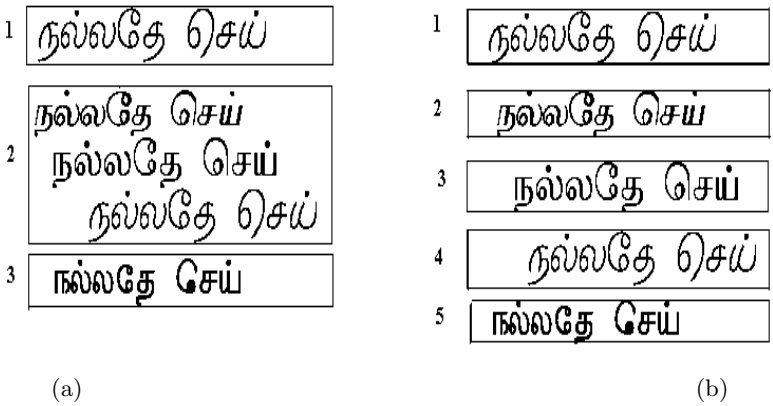


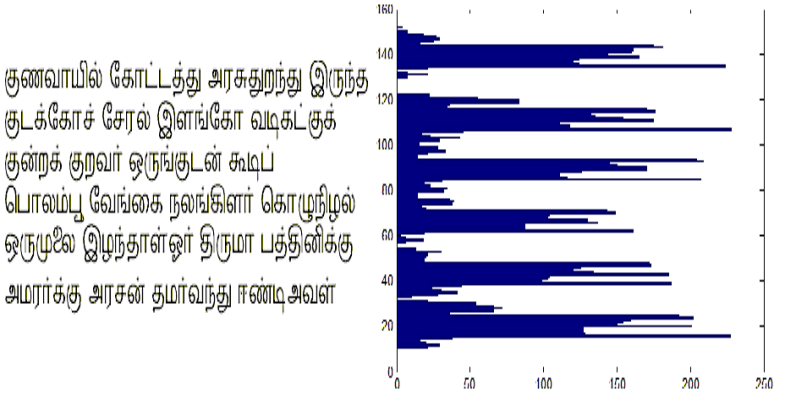
Fig. 6. (a) Various strips in Document 2 using Horizontal projection technique (b) Different lines identified using Proposed Method

4 rows from ending of a strip are ignored and the remaining rows in the strip are considered for processing.

A row with minimum horizontal projection is identified and based on that the strip is splitted into two strips. The value 4 is selected heuristically for this experiment image. This parameter value 4 need not be constant for all images. It may be differed based on the font size of the text considered for segmentation. The proposed method is implemented for different types of images with tamil characters. The experiment has been carried out with 50 different documents and the results for first three documents are discussed above and the results of first ten documents are listed in Table 1. It has been observed that every selected image is splitted 100% accurately.

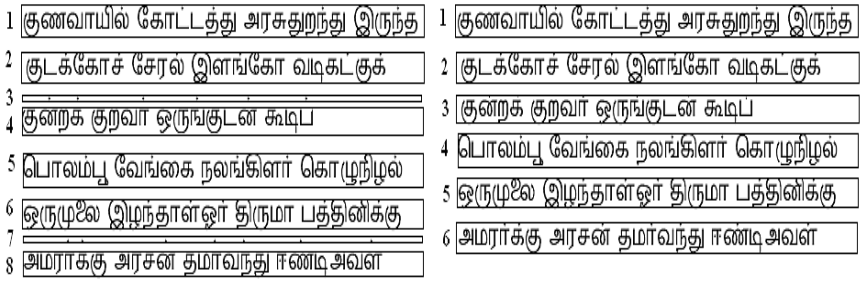
Table 1. Results of Line Segmentation by proposed

Document	# of lines	# of Strips : # of Overlapped lines	# of Normal lines	# of over segmented lines	# of segmented lines	Accuracy
Document1	13	2 : 3,2	7	1	13	100%
Document2	5	1 : 3	2	0	5	100%
Document3	6	0 : 0	4	2	6	100%
Document4	18	3 : 3,3,3	6	3	18	100%
Document5	17	3 : 5,2,3	7	0	17	100%
Document6	13	1 : 5	7	1	13	100%
Document7	15	0 : 0	12	3	15	100%
Document8	20	2 : 5,3	10	2	20	100%
Document9	12	3 : 2,2,2	6	0	12	100%
Document10	21	0 : 0	18	3	21	100%



(a) (b)

Fig. 7. (a) Original image for Document 3 and (b) Histogram of Document 3



(a) (b)

Fig. 8. (a) Text strips in Document 3 using Horizontal projection technique (b) Different lines identified using Proposed Method

5 Conclusion

The results of the experiment show that the proposed algorithm segments the overlapping lines into individual lines and joints the over segmented lines with the body of the text properly. All images considered for experiment are successfully segmented by the proposed algorithm. This algorithm is not using meanline and baseline concept. The steps involved in line segmentation method are very easy for implementation and all the lines are segmented properly eventhough the lines are touched.

References

1. Dhanya, D., Ramakrishnan, A.G., Pati, P.B.: Script Identification in printed bilingual documents. Sadhana 27, 73–82 (2002)

2. Jindal, M.K., Lehal, G.S., Sharma, R.K.: Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script. World Academy of Science, Engineering and Technology 21, 1153–1162 (2008)
3. Jindal, M.K., Sharma, R.K., Lehal, G.S.: Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts. International Journal of Computational Intelligence Research 3, 277–286 (2007)
4. Karthik, S., Hemanth, V.K., Balaji, V., Soman, K.P.: Level Set Methodology for Tamil Document Image Binarization and Segmentation. International Journal of Computer Applications 39, 7–12 (2012)
5. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line Segmentation of Historical Documents: A Survey. International Journal on Document Analysis and Recognition (2006)
6. Manmatha, R., Rothfeder, J.L.: A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents. IEEE Transactions on Pattern Analysis And Machine Intelligence 27, 1212–1225 (2005)
7. Stamatopoulos, N., Gatos, B., Perantonis, S.J.: A method for combining complementary techniques for document image segmentation. Pattern Recognition 42, 3158–3168 (2009)
8. Premaratne, H.L., Bigun, J.: A segmentation-free approach to recognise printed Sinhala script using linear symmetry. Pattern Recognition 37, 2081–2089 (2004)
9. Soujanya, P., Koppula, V.K., Gaddam, K., Sruthi, P.: Comparative Study of Text Line Segmentation Algorithms on Low Quality Documents. Special Issue of International Journal of Computer Science & Informatics (IJCSI) II(1,2), 2231–5292, ISSN (Print) : 2231-5292
10. Kunte, S., Samuel, S.: Two Stage Character Segmentation Technique for printed Kannada Text. Special Issue on Image Sampling and Segmentation (March 2006)
11. Siromony, G., Chandrasekaran, R., Chandrasekaran, M.: Computer Recogniton of printed Tamil Characters. Pattern Recognition 10, 243–247 (1978)
12. Sridevi, N., Subashini, P.: Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques. International Journal of Computer Applications 52, 7–12 (2012)
13. Garain, U., Chaudhuri, B.B.: Indian Script character recognition: a survey. Pattern Recognition 37, 1887–1899 (2004)
14. Garain, U., Chaudhuri, B.B.: Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis. IEEE Transaction on Systems, Man and Cybernetics-Part C: Application and Reviews 32, 449–459 (2002)
15. Garain, U., Sinha, S., Chaudhuri, B.B.: Multi-Script Line identification from Indian Documents. In: Proceedings of the ICDAR (ICDAR 2003), pp. 880–884 (2003)
16. Kumar, V., Sengar, P.K.: Segmentation of Printed Text in Devanagari Script and Gurmukhi Script. International Journal of Computer Applications 3, 24–29 (2010)
17. Dongre, V.J., Mankar, V.H.: Devnagari Document Segmentation Using Histogram Approach. International Journal of Computer Applications 1, 46–53 (2011)

eCS: Enhanced Character Segmentation – A Structural Approach for Handwritten Kannada Scripts

C. Naveena[1], V.N. Manjunath Aradhya[2], and S.K. Niranjan[2]

[1] Dept. of CSE, HKBK College of Engineering, Bangalore, India
[2] Dept. of Master of Computer Applications,
Sri Jayachamarajendra College of Engineering,
Mysore - 570 006, India
naveena.cse@gmail.com, aradhya1980@yahoo.co.in
sriniranjan@yahoo.com

Abstract. To build an efficient OCR system, preprocessing task of segmentation process should be in accurate way. In segmentation process, character segmentation plays an important role to obtain clear isolated characters. Character segmentation in Kannada word is a crucial task due to the presence of bottom extension characters (called as Vatthus in Kannada and as extra modifiers in English) and Modifiers. Due to the presence of modifiers and few cursive form of characters the script becomes semi-cursive while writing. With this nature some of the letters are touching each other and also bottom extension characters may get touch to main characters. In this regard, an enhanced Character Segmentation (eCS) approach is proposed for an unconstrained handwritten Kannada scripts. The method is based on thinning, branch point and mixture models. The Expectation-Maximization (EM) algorithm is used to learn the mixture of Gaussians. A cluster mean points are used to estimate the direction and branch point as a reference point for segmenting characters. We experimentally evaluated the proposed method on Kannada words and shown encouraging results.

Keywords: Character Segmentation, Thinning, Branch Points, Mixture Models, Kannada Script.

1 Introduction

Character segmentation has long been a critical area of the OCR process. It is important because incorrectly segmented characters are less likely to be recognized correctly. An OCR system may be designed to work for either on-line or on-line purposes. On-line OCR systems collect input data by recording the order of strokes written on an electronic bit-pad. On-line OCR systems do the same by recording pixel by pixel digital image of the entire writing with a digital scanner. OCR has a wide field of applications covering handwritten document transcription, automatic mail address recognition, machine processing of bank checks, faxes etc [1].

Segmenting characters from an unconstrained handwritten text is a difficult task because: (i) two consecutive characters of a word may touch (ii) two side-by-side

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 289–298, 2013.
© Springer International Publishing Switzerland 2013

non-touching characters are rarely vertically separable (iii) varies from individual to individual . In [2] basic segmentation algorithms are classified into three main categories: region, contour and recognition based methods. Zhao et al. [17], proposed an improved algorithm for segmenting and recognizing connected handwritten characters. The method gradient descent mechanism is used to weight the distance measure in applying KNN for segmenting/recognizing connected characters in the left to right direction. Tan et al. [13], presented handwritten character segmentation method based on nonlinear clustering. Two stage segmentation of unconstrained handwritten Chinese characters is reported in [16]. Maragoudakis et al [6], describes improved handwritten character segmentation by incorporating Bayesian knowledge with supprot vector machines. Zheng et al. [18], presented character segmentation system based on C# design and implementation. Sari et al. [11], presented on-line handwritten Arabic character segmentation algorithm based on morphological rules. Lee and Verma [5], presented binary segmentation algorithm for English cursive handwriting recognition. The binary segmentation algorithm is a hybrid segmentation technique and consists of over-segmentation and validation modules. The main advantage of binary segmentation technique is that, it adopts an unordered segmentation strategy.

Basu et al [1], presented segmentation of on-line handwritten Bengali script. In this method an isolated words are subdivided into four horizontal imaginary regions to segment a character. Selection of these regions is based on the basic characteristics of Bengali script. Pal et al [9], proposed touching numeral segmentation using water reservoir concept. A water reservoir concept is illustrated to find the touching regions in the numerals. A Reservoir is obtained by considering accumulation of water poured from the top or from the bottom of the numerals. Based on the analysis of reservoir boundary, touching position is detected. Sharma and Singh [12], proposed segmentation of handwritten text in Gurumukhi script. The segmentation of half characters in handwritten Hindi text is presented in [4].

From the above literature, many methods on handwritten character segmentation have been reported in English, Chinese and Arabic scripts. Also some works are carried out in Indian scripts such as Bengali, Gurumukhi. Recently, work on character segmentation for online Kannada text can be seen in [7, 10]. To the best of our knowledge, it is of first kind in the literature for an online handwritten Kannada character segmentation.

The outline of the paper is as follows: In section 2, properties of Kannada script are explained. In section 3, the proposed methodology is detailed. Experimental result is presented in section 4. Finally, conclusion is drawn.

2 Properties of Kannada Script

Kannada script is written horizontally from left to right and an absence of lower and upper case like in English language. Moreover, the Kannada characters are formed by combination of basic symbols, segmentation of the Kannada character is complex and challenging task & increased character set, it contains Vowels, Consonants & Compound characters. Some of the character may get overlap together. Kannada text is difficult when compared with Latin based languages because of its structured

complexity. Moreover, Kannada language uses 49 phonemic letters and it is divided into 3-groups,Vowels (Swaragalu- Anusvara (o), & Visarga (:)15), Consonants (Vyanjanagalu-34) and modifier glyphs (Half-letter) from the 15 vowels are used, to alter the 34 base consonants, creating a total of (34*15) +34=544 characters, sample of modifier glyphs additionally a consonants emphasis glyph called Consonant conjuncts in Kannada (vattakshar/ also called extra modifiers in English), exists for each of the 34 consonants. This gives total of (544*34) +15=18511 distinct characters [14].

3 Proposed Method

This section presents the enhanced Character Segmentation approach to the earlier work presented in [8]. Here in this work, segmentation-then-recognition approach is used to enhance the segmentation process. Initially, all components in a word image are detected by connected component analysis (CCA) algorithm which is as shown in the Figure 1. For a component c_i, its height and width are represented by h_i and w_i respectively. From this process the components those having average and below average height and width components are segmented, the remaining components are considered as touching components. To segment these touching components, we follow three steps namely: Thinning, Branch Points and Mixture Models. The steps are explained in following subsections.

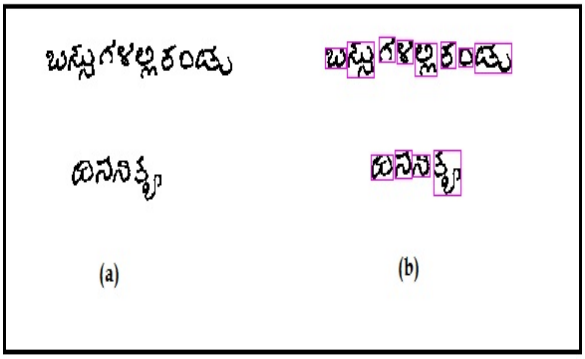


Fig. 1. (a)Input images (b) Results obtained after applying CCA algorithm

3.1 Thinning and Branch Points

In Kannada, some of the letters are cursive in nature and the script become semi-cursive while writing. Also from the statistical analysis most of the touching portion can be find within the half height from the bottom of character. These touching causes mainly because of modifiers and extra modifiers. To find this touching portion in the components, we have applied morphological thinning operation to touching components for further process, which is as shown in the Figure 2. Then, the thinned image is used to find the touching portion using branch points from templates of branch point identifiers. A branch point is a junction point, it connects three branches like a

capital T rotated by different angles. In this work, we have used 16 different branches of templates and each one has three branches, Figure 3 shows the several occurrences of 16 types. Figure 4 shows the branch points present in the thinned image and selection of best branch point from many points for accurate segmentation is the main task. For this purpose, we statistically analyzed that most of the touching portions are present at the right half of the average width of a component. Hence, we choose right most branch point as a segmentation point of the touching components. After getting the segmentation point, it is not straight forward way to cut touching portion like horizontal or vertical direction. This is due to the circular nature of Kannada script. To resolve this issue, we applied Mixture Models to thinned image to find the angle of direction to obtain an accurate segment of touching characters.

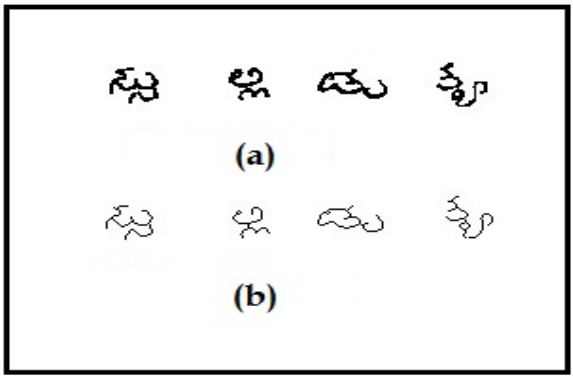


Fig. 2. Results of morphological thinning operation: (a) input image (b) thinned image

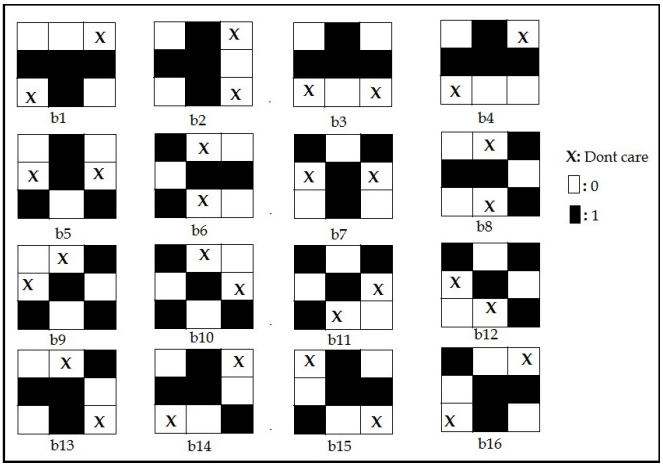


Fig. 3. Branch points extracting templates

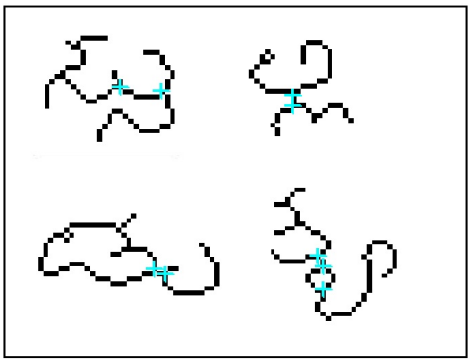


Fig. 4. Results obtained after applying branch point templates

3.2 Mixture Models

Mixture models use a set of data points as a input for clustering input data. Finding a clusters in a set of data points is a considerable problem. To obtain marginalization from joint distribution over observed and latent variables is relatively complex. To solve this problem, use of latent variable in a mixture distributions in which the discrete latent variables can be interpreted as defining assignment of data points to specific components of the mixture. Expectation-Maximization (EM) algorithm is one of the technique for finding maximum likelihood estimation in latent variable. More detailed description regarding EM algorithm can be seen in [3]. The advantage of using EM algorithm is that, it generalizes more complex models with combinations of discrete and real valued hidden variables. Also, it provides grouped or clustered observation even though for incomplete data and missing information.

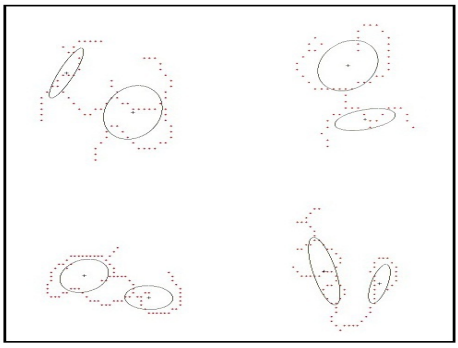


Fig. 5. Results obtained after EM algorithm

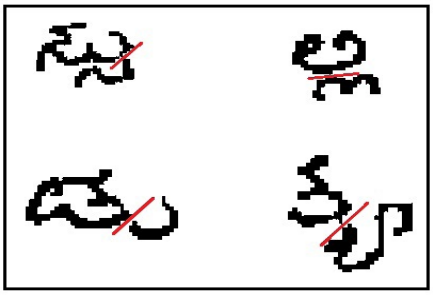


Fig. 6. Segment the touching character based on the skew angle orientation (solid line represents the skew angle orientation)

Means of two clusters are used for estimating the skew of a segmented touching component, because touching characters normally contain two components. Therefore in this work, two clusters are enough to find a skew angle. Figure 5 depicts the mean

points obtained for the input skewed word using mixture-of-gaussians. In this work, we have used skew estimators as Linear Regression Analysis (LRA) to estimate the skew angle of the touching character. Finally, with the reference to skew angle direction and best branch point, we segment the touching character which is as shown in the Figure 6.

4 Experimental Results

This section presents the results obtained in the conduct of experiment to study the performance of the proposed method. The method has been implemented in MATLAB 10.0 on a Core2duo processor with 1GB RAM. For our experiment we have collected a data set comprising 400 handwritten Kannada words. Most of the words are presented with one or two touching components. Figure 7 shows the samples of word images of our own collected dataset. To find the segmentation accuracy there are two fundamental analytical segmentation strategies, which are segmentation-then-recognition and segmentation-base strategy.

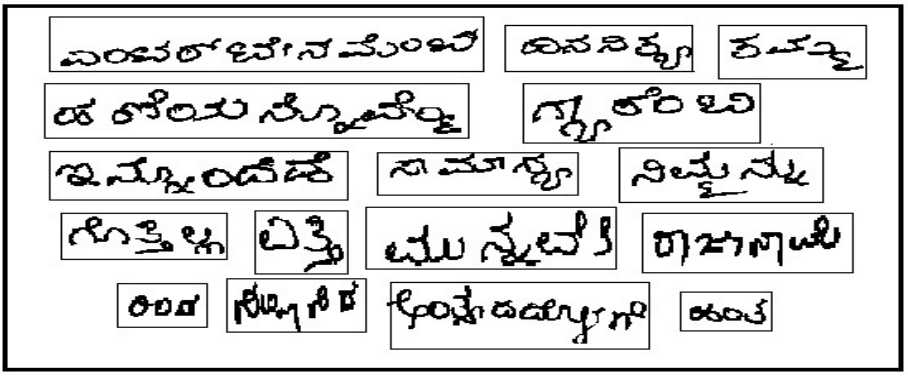


Fig. 7. Sample images of Kannada words

The segmentation accuracy measure as follows:

$$Segmentation\ Accuracy = (SC/N) * 100 \quad (1)$$

where *SC* is a complete segmented characters in the dataset. *N* is a total number of characters presented in the dataset.

Initially, we adopted a segmentation-base strategy, in that explicit segmentation is used to segment a word into an ideal part of isolated characters. In this strategy, the proposed method achieves a segmentation accuracy of 85.5%. In this, we choose right most branch point as a segmentation point. But in few cases, this assumption leads to an incorrect segmentation, which is shown in Figure 8. To resolve this, we used segmentation-then-recognition strategy. In this, we first trained priory segmented characters of modifiers and extra modifiers using PCA algorithm [15]. Initially, touching

characters are segmented by right most branch point and these segmented characters are subjected to recognition process with trained characters. A Euclidean distance measure is used as a classifier in recognition process.

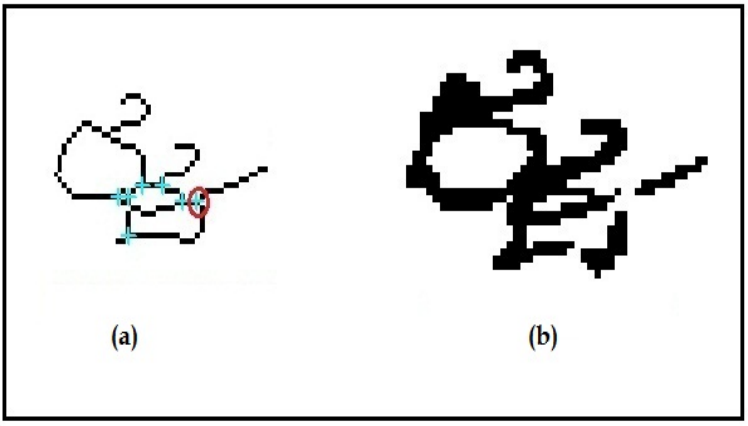


Fig. 8. Incorrect segmentation by selecting right most branch point as segmentation point: (a) Selecting a right most branch point (b) Incorrect segmentation.

Suppose, if the segmented characters are not recognized correctly again segmentation needs to be performed like a feedback process, which is based on a another branch point nearer to right most branch point. Then after recognition process is performed. This process is repeated until the best branch point segments character correctly. Figure 9, shows correctly recognized segmented characters. By adapting segmentation-then-recognition strategy, the proposed method segmentation accuracy increases around 4%. The successful segmentation results are shown in Figure 10. The proposed method fails in two issues. First, if the branch points are not encountered in touching portion, which is shown in Figure 11. Second, if more than two components get touched/overlapped to each other, which is shown in Figure 12.

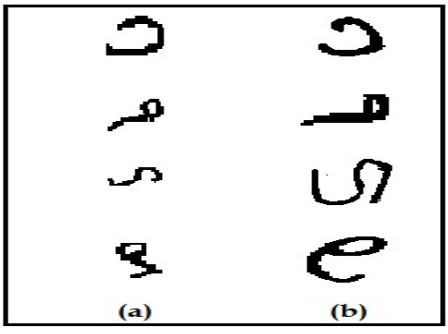


Fig. 9. Recognized the characters after segmentation: (a) segmented characters (b) correctly recognized characters

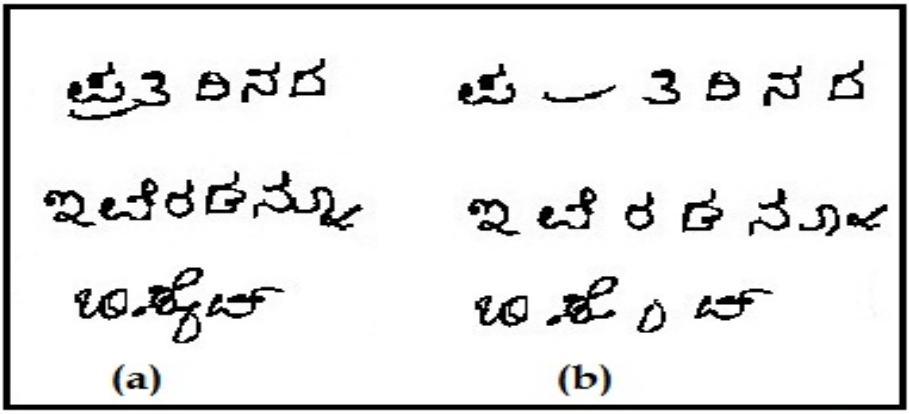


Fig. 10. Successful results obtained for the proposed method: (a) input images (b) segmentation results

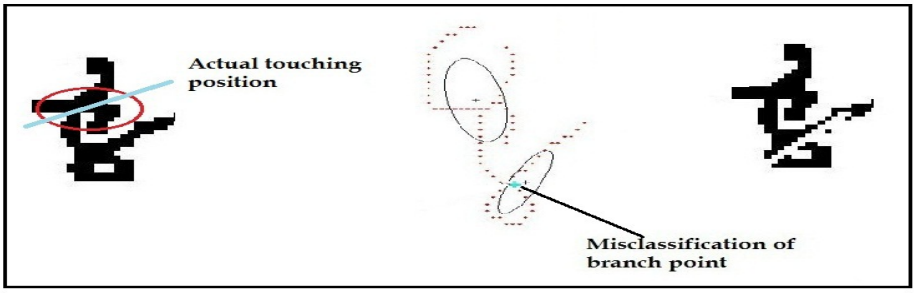


Fig. 11. Failure case of the proposed method: branch points are not present in the touching portion of the component

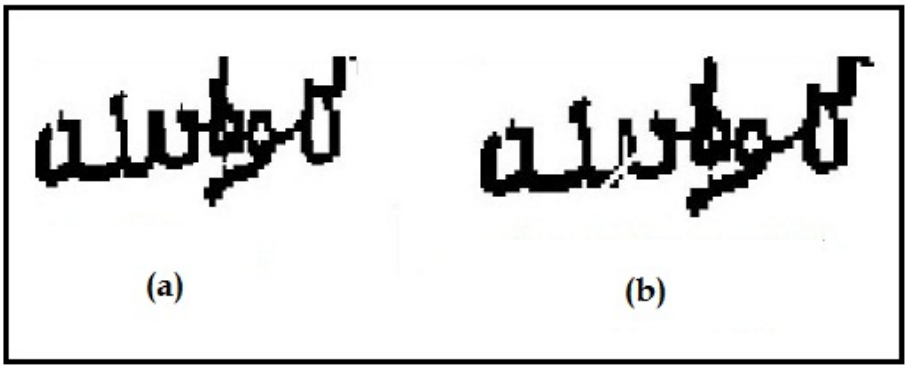


Fig. 12. Failure case of the proposed method: more than two components touching each other, (a) input image (b) incorrect segmentation

5 Conclusion

In this paper, enhanced character segmentation of Kannada script is presented. The proposed method is based on the thinning, branch point and mixture models. The thinning is the most commonly adopted technique to skeletonize an input image and is used to find the branch points present in an image. From the reference of the best branch point and skew angle obtained from the mixture models, we segment the character from touching character. The proposed method is tested on handwritten Kannada words and experimentation is performed based on two strategies: segmentation-base strategy and segmentation-then-recognition strategy. In segmentation-based experiment the selected best branch point may leads to incorrect segmentation in some touching characters and it will reduces the segmentation accuracy. To resolve this issue, we use segmentation-then-recognition strategy. With this strategy, we selected the best branch point and the system's segmentation accuracy increased around 4%. In future, we plan to solve failure issues to increase the segmentation accuracy.

References

1. Basu, S., Chaudhuri, C., Kundu, M., Nasipuri, M., Basu, D.K.: Segmentation of online handwritten Bengali script. In: Proceedings of 28th IEEE ACE, pp. 171–174 (2002)
2. Bhowmik, T.K., Roy, A., Roy, U.: Character segmentation for handwritten Bangla words using artificial neural network. In: Proceedings of 1st IAPR TC3NNLDAR (2005)
3. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
4. Garg, N.K., Kaur, L., Jindal, M.K.: The Segmentation of Half Characters in Handwritten Hindi Text. In: Singh, C., Singh Lehal, G., Sengupta, J., Sharma, D.V., Goyal, V. (eds.) ICISIL 2011. Communications in Computer and Information Science, vol. 139, pp. 48–53. Springer, Heidelberg (2011)
5. Lee, H., Verma, B.: Binary segmentation algorithm for english cursive handwriting recognition. Pattern Recognition 45(4), 1306–1317 (2012)
6. Maragoudakis, M., Kavallieratou, E., Fakotakis, N.: Improving handwritten character segmentation by incorporating Bayesian knowledge with support vector machines. In: Proceedings of ICASSP 2002, vol. 4, pp. IV-4174 (2002)
7. Mohan, P., Shashikiran, K., Ramakrishnan, A.G.: Unrestricted kannada online handwritten akshara recognition using sdtw. In: ACM - Proceedings of International Workshop on Multilingual OCR (2009)
8. Naveena, C., Manjunath Aradhya, V.N.: Handwritten character segmentation for kannada scripts. In: Proceedings of 2nd World Congress on Information and Communication Technologies (WICT 2012), pp. 144–149 (2012)
9. Pal, U., Belaid, A., Choisy, C.: Touching numeral segmentation using water reservoir concept. Pattern Recognition Letters 24(3), 261–272 (2003)
10. Prasad, M.M., Sukumar, M., Ramakrishnan, A.G.: Divide and conquer technique in online handwritten kannada character recognition. In: Proceedings of International Conference on Signal Processing and Communications (SPCOM), pp. 1–5 (2010)
11. Sari, T., Souibi, L., Sellami, M.: Off-line handwritten Arabic character segmentation algorithm: ACSA. In: Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2002), pp. 452–457 (2002)

12. Sharama, R.K., Singh, A.: Segmentation of handwritten text in Gurumukhi script. International Journal of Image Processing 2(3), 12–17 (2004)
13. Tan, J., Lai, J.H., Wang, C.D., Wang, W.X., Zuo, X.X.: A new handwritten character segmentation method based on nonlinear clustering. Neurocomputing 89(15), 213–219 (2012)
14. Thungamani, M., Kumar, P.R.: A survey of methods and strategies in handwritten kannada character segmentation. International Journal of Science Research 1(1), 18–23 (2012)
15. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
16. Zhao, S., Chi, Z., Shi, P., Yan, H.: Two-stage segmentation of unconstrained handwritten Chinese characters. Pattern Recognition 36(1), 145–156 (2003)
17. Zhao, X., Chi, Z., Feng, D.: An improved algorithm for segmenting and recognizing connected handwritten characters. In: Proceedings of 11th International Conference on Control, Automation, Robotics and Vision, pp. 1611–1615 (2010)
18. Zheng, Z., Zhao, J., Guo, H., Yang, L., Yu, X., Fang, W.: Character segmentation system based on C# design and implementation. In: Proceedings of 2012 International Workshop on Information and Electronics Engineering (IWIEE). Procedia Engineering, pp. 4073–4078. Elsevier (2012)

Image Restoration by Using Evolutionary Technique to Denoise Gaussian and Impulse Noise

Nallaperumal Krishnan[1], Subramanyam Muthukumar[1], Subban Ravi[2], D. Shashikala[2], and P. Pasupathi[2]

[1] Centre for IT and Engg., Manonmaniam Sundaranar University, Tirunelveli, India
[2] Dept. of Computer Science, Pondicherry University, Pondicherry, India
krishnann@computer.org, {su.muthukumar,sravicite}@gmail.com

Abstract. Most of the techniques for image restoration are based on some known degradation models. Here a genetic algorithm based filter is used to restore the degraded image without having any prior knowledge about the blurring model or noise type. The observed degraded image is denoised and the initial target image is generated by blind deconvolution technique using higher-order statistics. Recombination and mutation mechanisms are implemented to create better individuals. More good solutions are generated by the selection of fittest individuals. The selection procedure is based on the similarity of the individuals with the target image. In this method, the initial target image is obtained by significantly removing noise with both Gaussian and non-Gaussian probability distributions, hence the convergence of the solution set becomes faster.

Keywords: Blind Deconvolution, Color Image Restoration, Higher Order Statistics.

1 Introduction

The field of image restoration is very broad with an over 30 year long history. Since it has many applications and has a simple mathematical formulation it has attracted strong research interest. It is a classical inverse problem for which good solutions are not easily obtained. The objective of image restoration [1] is to estimate an image from a given corrupted version, using an inverse operation like a linear filter. Optimal linear filter known as minimum mean square error filter or nonlinear filter are very useful in image restoration from corrupted images. Images may be corrupted by both positive and negative impulse noise [2]. Non linear mean filter cannot remove such positive and negative impulse noise simultaneously but the median filter performs quite well in such cases. But as the probability of impulse noise occurrence becomes high it fails. To overcome this situation two new algorithms for adaptive median filters are proposed in [3]. Prior information is used in usual restoration approaches to restrict the number of possible solutions. Such prior knowledge can be stochastic or deterministic in nature. Learning-based algorithms for image restoration and blind image restoration are proposed in [4], which addresses linear degradation systems which are spatially invariant, and has typically low-pass characteristics. Genetic algorithms are most powerful optimization technique in large solution space, for a natural selection of fittest individuals.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 299–309, 2013.
© Springer International Publishing Switzerland 2013

Efficiency of genetic algorithms [5] depends on the selection process and exchange of genetic material. Various fitness functions are used for the selection process and cross-over and mutation operations are used to construct next generation images. The genetic algorithms produce better results in various specialized applications with faster processing times. Genetic algorithms are used in various fields such as image enhancement, segmentation, feature extraction and classification as well as the image generation. As the most powerful optimization technique, genetic algorithms are getting increasing popularity [6]. In genetic algorithms the statistics about the search space is maintained implicitly by a population of potential solutions to a given problem. Color or structural characteristics of the traditional color filtering scheme can be improved using genetic algorithm based filter optimization technique. Acceptable noise attenuation capabilities are exhibited by Optimized filters. Various tasks from basic image contrast and level of detail enhancement, to complex filters and deformable models parameters are solved using this paradigm. Since it consider a large number of possible solutions, the algorithm allows to perform robust search without trapping in local extremes. This also makes genetic algorithms capable of solving a very big variety of simple and difficult tasks [7].

The idea of improving the performance of genetic algorithms using implicit statistics is explored in [8], which proposes a statistics-based adaptive non-uniform crossover. Genetic algorithm based image restoration problem is discussed in [9]. The genetic algorithm is used to restore the image from the initial target image, without having any prior information about the degradation model and noise type. Usual approach in image restoration is to explicitly specify the numerical value of the regularization parameter. The initial target image is generated using a Weighted Order Statistics filter, which is used as the reference image and degraded images by unknown models are restored in successive stages. Using the WOS the regularization parameter is adjusted in successive stages. The quality of the images seems to improve in successive stages.

2 Methodology

An evolution theory based image restoration technique is proposed here, which is a modification to the genetic algorithm based filter using Weighted Order Statistics described in [9]. The selection process is based on the reference image which is generated using Weighted Order Statistics, and the results obtained are not much satisfactory. Solution set can be improved if the observed degraded image is denoised [10] and the initial target image is generated by blind deconvolution technique using higher-order statistics.

2.1 Evolutionary Neuro Fuzzy Models

Evolutionary neuro fuzzy (ENF) system is the consequence of adding evolutionary search practices to the system, integrating fuzzy logic computing and neural learning. With these techniques the limitations of the existing hybrid systems can be overcome. The main objective and the drawback of the NFS is that, the learning techniques is

based on the gradient descent optimization technique [11]. That is, in back propagation, training will not coverage and tuning of the membership function parameters through neural learning is not guaranteed. The algorithm will be trapped in local minima. With this kind of system, the global solution is impossible to find.

Genetic Algorithms (GA) also known as "Evolutionary Algorithms" are population based algorithms. Evolutionary algorithms are iterative probabilistic algorithms, which are used to in real time problems to optimize their solutions [12]. Populations of individuals represented as: $p(t) = \{p_1^t, \dots, p_n^t\}$ which is designed for all iterations of t. The members of the populations are evaluated based on fitness function. A new set of population is evolved at $(t+1)^{th}$iteration based on robustness of the gene, based on mutation or crossover operations. In genetic algorithms (GA), the populations are coded as chromosomes which are the binary strings.

Evolutionary Algorithms (EA) is inspired by Darwins' idea of 'survival of the fittest' [13], which is the element that forms the chromosome. EAs are comparable with GAs. The distinctive feature of the EA algorithm is that, all the chromosomes and their off springs are allowed to gain adequate experience, by a process of local searching, before involving in the evolutionary process. The first population created is random which is similar to the population generation procedure in GA. Later, a global search is performed on each individual member to improvise its experience and thus obtain a population of local optimum solutions. Then, crossover and mutation operations are performed to reproduce the off springs [14]. These off springs are then subjected to the local search so that local optimality always maintained in the system.

In this paper, a novel method using Evolutionary Algorithms based neuro-fuzzy is proposed to harness the power of fuzzy reasoning and the learning capabilities of neural networks [11].

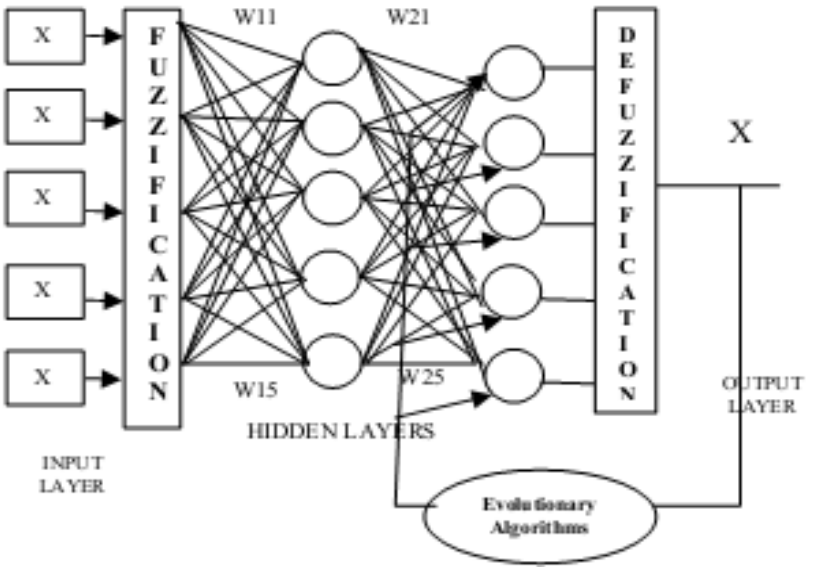


Fig. 1. Evolutionary Model

2.2 Higher Order Statistics

Higher Order Statistics (HOS) [15] measures are extensions of second-order measures to higher orders. The second-order measures work fine if the signal has a Gaussian probability density function. Any Gaussian signal is completely characterized by its mean and variance. Consequently the HOS of Gaussian signals are either zero or contain redundant information. Many signals encountered in practice have non-zero HOS, and many measurement noises are Gaussian, and so in principle the HOS are less affected by Gaussian background noise than the second order measures [16].

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. A negative skew indicates that the left side of probability density function is longer than the right side and the bulk of the values including the median lie to the right of the mean [17]. A positive skew indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean [18]. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution.

For a sample of *n* values the sample skewness is

$$G_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}} \quad (1)$$

where, $\bar{x}$ is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance. The normal distribution has a skewness of zero and a nonzero skewness of the dataset indicates whether deviations from the mean are going to be positive or negative [22].

Kurtosis is a measure of the peakedness of the probability distribution of a real-valued random variable [19]. Higher kurtosis means more of the variance is the result of infrequent extreme deviations. Kurtosis is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3. For a sample of *n* values the sample kurtosis is

$$G_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} - 3 \quad (2)$$

where, m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean or the sample variance, x_i is the i^{th} value, and $\bar{x}$ is the sample mean. Noise pixels are separated by computing the skewness and Kurtosis of random samples around the pixels, and then a selective median filtering is applied.

After generating the initial target image initial solution set is generated using the following sequence of operations: Thresholding, Convolution, Color Addition and Median filtering.

The idea here is to perform various operations in order to improve the image quality. Select the fit individuals to generate more and more good solutions. The solution set ultimately convergence to sufficiently high point of fitness. Recombination and mutation mechanisms are implemented to create better individuals [20].

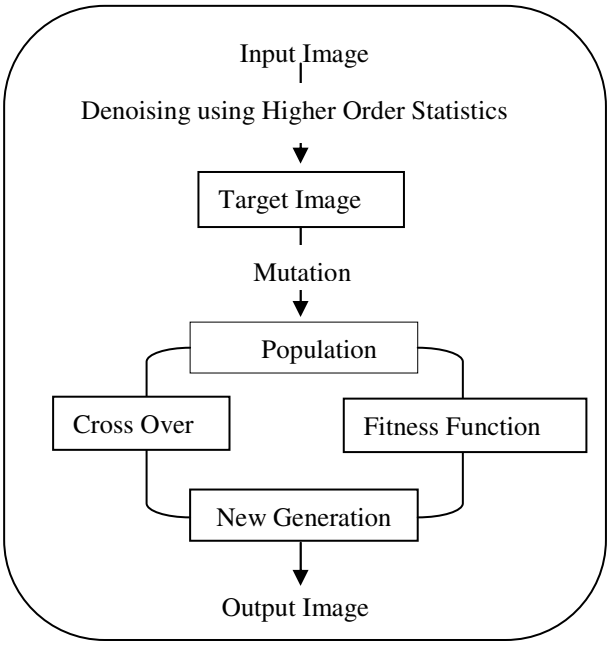


Fig. 2. Algorithm Diagram

A. Fitness Function

The selection procedure is based on the similarity of the individuals with the target image. There are various methods to measure the similarity between the images. The best measure is the human subjective judgment. Popular measures of performance for evaluating the different between the original and filtered images includes [10],

i. Peak Signal to Noise Ratio (PSNR) and

ii. Mean Different per Pixel (MDPP)

In many applications the error is expressed in terms of a signal to noise ratio (SNR) and is given as equation 3.

$$SNR = 10\log_{10}\frac{\sigma^2}{MSE}dB \tag{3}$$

where σ^2 is the variance of the desired or original image. The peak signal to noise ratio (PSNR) is expressed as equation 4.

$$PSNR = 10\log_{10}\frac{255^2}{MSE}dB \tag{4}$$

The fitness function using MDPP is given by

$$\text{MDPP} = \frac{1}{\text{N}\times\text{X}}\sum_{\text{i,j}=1}^{\text{N}}|\text{orig(i,j)}.\,\text{filt(i,j)}| \tag{5}$$

The general features of EA which make it suitable for the likelihood of choosing the operator is given in figure 2. The initial population which is chosen by the EA

algorithms is the one individual string which is defined as the 'queen' string. The generations are generated by conducting mutation operations on these strings. These strings contain the membership function (mf) width parameter P and threshold weights have to be applied between the inputs and hidden layers [21]. The descriptions of the steps of the evolutionary algorithm are discussed as:

1. Randomly generate the individual string (queen string).
2. Initialize the queen string to the bit string of 0's and 1's.
3. Child string is generated using the mutation operator.
4. Then, the mutation points are chosen at random.
5. For each string, perform the following:
    a. Interpret the parameter values.
    b. Assign to the neuro-fuzzy filter.
6. Stop the process after 100 generations.

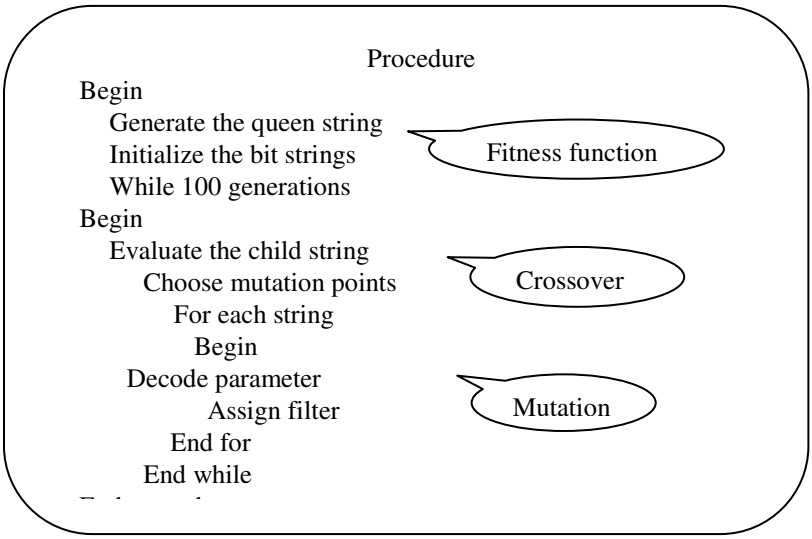


Fig. 3. EA for parameters optimization

B. Selection Method

The individuals are arranged in the descending order of the probability of that individual being selected. Based on the limit of solution set size, individuals with high rank in the list are selected. A simple linked list can be used to keep track of the ordering of the individuals.

C. Crossover Operator

New generation of individuals are generated using crossover operators. A crossover operator takes two individuals called parents and combines them to make the offspring. Various mathematical operations such as average, min, max etc. can be used in crossover operations. Uniform or random pixels can be considered in crossover operation.

D. Image Similarity Measure

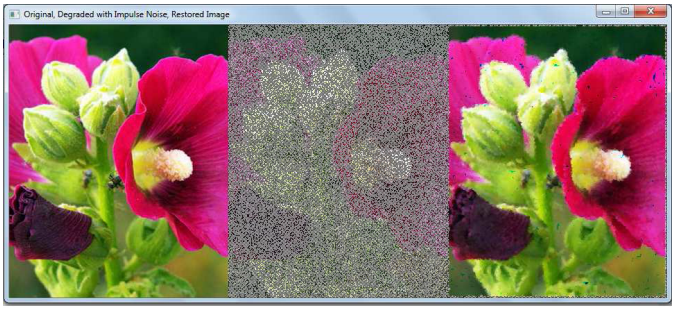
Image entropy and histogram can be used to measure the similarity between images. These measures are used to compare the images. The human observer can control the selection of better individuals based on these measures.

Table 1. The GA configuration of test images

Operators	**Values**
Initial pool entered	Randomly
Population size	20-30 Individuals
Chromosome length	5
Selection operator	Roulette wheel
Crossover operators	CX (single-point)
Mutation operator	Randomly
Crossover rate rang	0.60 - 0.90
Mutation rate range	0.0011- 0.0033
Maximum generations range	40 - 50
GA mask	(R U(0-1)

3 Results and Discussion

The algorithm is implemented in IDL. In order to test the algorithm, the original images are first degraded by applying impulse noise, Gaussian noise and the combination of both. The results obtained are shown in the next page.

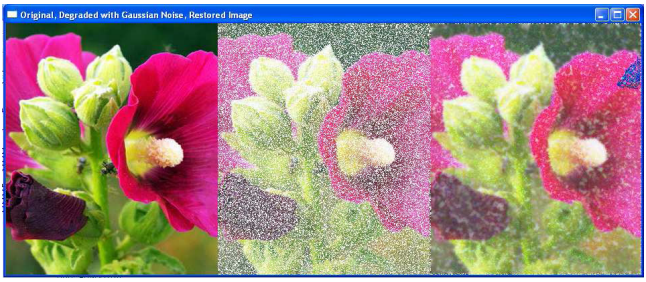


(a) (b) (c)

Fig. 4. (a) Original (b) Degraded with 90% impulse noise (c) restored image

(a) (b) (c)

Fig. 5. (a) Original (b) Degraded by Gaussian noise with variance 50 (c) Restored Image



(a) (b) (c)

Fig. 6. (a) Original (b) Image degraded by both impulse and Gaussian noise (c) Restored Image

Table 2. PSNR for impulse noise varying from 10 to 90 Percentages

Noise %	Impulse_PSNR (EA)	Gaussian_PSNR (EA)
10	34.1287	32.8667
20	33.9478	31.9127
30	32.8472	30.8053
40	32.7344	30.0004
50	32.5957	29.1287
60	30.6309	28.3019
70	29.5560	28.1150
80	28.7829	28.5581
90	25.7834	27.6538

In evolution theory based algorithms recombination of the individuals gain much attention, as it extend the features of selected individuals to next generation and helps in finding good solutions. But the recombination solutions obtained are not locally optimal always. Thus it is desirable to locally optimize each cross over solutions before adding to the solution set. The parameterized uniform cross over may yield better

Table 3. Impulsive Noise Performance

	Noise%	MF	WMF	CWMF	AMF	EA
PSNR(db)	**10**	30.2572	31.0835	31.0437	34.1047	**35.4715**
	30	25.8437	26.3175	25.5848	32.5861	**33.2864**
	50	22.7853	26.5076	26.7352	31.6955	**31.8552**
	70	13.3704	20.6964	21.7738	21.4689	**29.0352**
	90	6.56926	14.6863	15.3842	9.14986	**18.0228**
IQI	**10**	0.7118	0.6734	0.6701	0.9277	**0.92772**
	30	0.5998	0.5787	0.5868	0.6694	**0.66938**
	50	0.3438	0.2594	0.2832	0.5686	**0.54747**
	70	0.0914	0.2594	0.2832	0.5686	**0.39662**
	90	0.0068	0.0415	0.0514	0.0133	**0.19914**

Table 4. Both impulse and Gaussian noise Performance

	Noise%	MF	WMF	CWMF	AMF	EA
PSNR(db)	**10**	19.008	18.11	18.04	26.134	**26.4346**
	30	18.307	16.81	17.01	22.615	**22.6150**
	50	16.426	14.86	15.13	19.355	**19.5416**
	70	11.485	12.47	12.70	15.554	**17.0378**
	90	6.3801	9.097	9.413	8.5209	**12.6441**
IQI	**10**	0.5998	0.561	0.555	0.9221	**0.92207**
	30	0.5712	0.484	0.484	0.8693	**0.86933**
	50	0.4784	0.374	0.385	0.7705	**0.75316**
	70	0.2233	0.245	0.253	0.5727	**0.56162**
	90	0.0049	0.009	0.098	0.1128	**0.32057**

results. Since the selection procedure is based on the initial target image, the generation of the initial target image is of at most importance in this approach. Better results can be achieved by preprocessing the degraded observed image using Nonlocal Image Averaging before applying the higher order statistics.

4 Conclusion

An evolution theory based image restoration technique is proposed in this paper. After the initial target image is obtained, evolution theory based technique is used to restore the image. Initial target image is generated by blind deconvolution technique using higher-order statistics; this initial target image can be used as a reference to select the fittest individuals. The human observer can control the selection process. Here only simple mathematical operations are considered for the cross over operation. Here the number of iteration depends on the PSNR. Better solution is to control the number of iterations according to the choice of a human observer. Better quality images can be

achieved after successive iterations. It can be done by maintaining the solution set in each successive generation in the descending order of the probability of that individual being selected.

References

[1] Athans, M., Tse, E.: A direct derivation of the optimal linear filter using the maximum principle. IEEE Transactions on Automatic Control 12(6), 690–698 (1967)
[2] Bernstein, R.: Adaptive nonlinear filters for simultaneous removal of different kinds of noise in images. IEEE Transactions on Circuits and Systems 34(11), 1275–1291 (1987)
[3] Hwang, H., Haddad, R.A.: Adaptive Median Filters: New Algorithms and Results. IEEE Transactions on Image Processing 4(4) (1995)
[4] Nakagaki, R.: A VQ-Based Blind Image Restoration Algorithm. IEEE Transactions on Image Processing 12(9) (September 2003)
[5] Gen, M., Cheng, R.: Genetic algorithms and engineering optimization, vol. 7. John Wiley & Sons (2000)
[6] Sivanandam, S.N., Deepa, S.N.: Introduction to genetic algorithms. Springer (2007)
[7] Whitley, D., Starkweather, T., Bogart, C.: Genetic algorithms and neural networks: Optimizing connections and connectivity. Parallel Computing 14(3), 347–361 (1990)
[8] Yang, S.: Adaptive Crossover in Genetic Algorithms Using Statistics Mechanism, Artificial Life
[9] Chickerur, S., Aswatha Kumar, M.: Biologically Inspired Filter For Image Restoration. Int. J. of Recent Trends in Engr. 2(2) (November 2009)
[10] Muthukumar, S., Raju, G.: A non-linear image denoising method for salt and pepper noise removal using Fuzzy-Based approach. In: 2011 International Conference on Image Information Processing (ICIIP). IEEE (2011)
[11] Khanmirzaei, Z.: Training Recurrent Neuro-Fuzzy System Using Two Novel Population-Based Algorithms for Temperature Forecasting. In: IEEE 10th International Conference on Computer and Information Technology (CIT). IEEE (2010)
[12] Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of global optimization 39(3), 459–471 (2007)
[13] Praveen Kumar, K., et al.: Memetic NSGA-a multi-objective genetic algorithm for classification of microarray data. In: International Conference on Advanced Computing and Communications, ADCOM 2007. IEEE (2007)
[14] Gerbex, S., Cherkaoui, R., Germond, A.J.: Optimal location of multi-type FACTS devices in a power system by means of genetic algorithms. IEEE Transactions on Power Systems 16(3), 537–544 (2001)
[15] Nikias, C.L., Mendel, J.M.: Signal processing with higher-order spectra. Signal Processing Magazine. IEEE Signal Processing Magazine 10(3), 10–37 (1993)
[16] NOZ, Antonio MORENO-MU for Characterization of Electrical Power Quality Signals using Higher Order Statistical Features. Technology 13: 14
[17] El-Shafie, A., Jaafer, O., Seyed, A.: Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia. Int. J. Physical Sci. 6(2), 2875–2888 (2011)
[18] Khazaei, H., Misic, J., Misic, V.B.: Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems. IEEE Transactions on Parallel and Distributed Systems 23(5), 936–943 (2012)

[19] Serfling, R.: Nonparametric multivariate descriptive measures based on spatial quantiles. Journal of Statistical Planning and Inference 123(2), 259–278 (2004)
[20] Deb, K., Goel, T.: Controlled elitist non-dominated sorting genetic algorithms for better convergence. In: Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., Corne, D.W. (eds.) EMO 2001. LNCS, vol. 1993, p. 67. Springer, Heidelberg (2001)
[21] Alamelumangai, N., Devishree, J.: A novel CAD system for breast cancer segmentation in sonograms (2006)
[22] Muthukumar, S., Pasupathi, P., Deepa, S., Krishnan, N.: An efficient color image denoising method for Gaussian and impulsive noises with blur removal. In: 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–4. IEEE (December 2010)

Digital Watermarking Using Modified Techniques in Spectral Domain of Images

Yashwanth Kanduri and Madhuri Midatala

National Institute of Technology,Warangal
yashwanth3030@live.com, madhurimidatala4141@gmail.com

Abstract. Digital watermarking using advanced techniques in spectral domain and pictorially showing the comparisons of importance and efficiency of discrete fourier transform and discrete cosine transform in spectral domain of digital watermarking is mainly reflected in this paper. The proposed technique will produce more efficient output but under certain specifications. The comparisons of various transforms are essentially performed and contemplated for better results .The features of the digital world lead to economical changes such as cheap distribution and also serious risks in simplifying unauthorized copying and distribution. So, there is an immediate requirement for a well-established and advanced technique for digital watermarking.

1 Introduction

Watermarking can be used in a wide variety of applications. In general, if it is useful to associate some additional information with it, this metadata can be embedded with the watermark [1] . A visible watermark is an opaque or semi transparent sub-image or image placed at the top of another image so that it is obvious to the viewer. An invisible watermark cannot be seen with the naked eye. They are imperceptible but can be recovered. The embedded information is hidden (in low value bits or least significant bits of the picture pixels, frequency or other value domains and linked in separably with the source data structure. Technically speaking, a code conveying some important information about the legal data owner, or the allowed uses of data is hidden within the data itself instead of being attached to the data as a header or separate file [1-3].

Nomenclature

LVSP – Linear valued spectral processed formula; NVSP – n-valued spectral processed formula; IVSP – Infinite valued spectral formula; ENVSP – Extended n-valued spectral formula; ENVSP – Extended infinite valued spectral formula.

Watermarking is classified as:

- Watermarking in spatial domain.
- Watermarking in spectral domain.

Watermarking in spatial domain is a method in which the secret message has to be embedded in the asset. So, the technique implemented here is the asset is broken

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 310–315, 2013.
© Springer International Publishing Switzerland 2013

down to 8-bit planes and the most significant bits of message are embedded in the place of the least significant bits of the asset [3,4,7]. Watermarking in spectral domain is a method in which image is brought to frequency domain by using variety of transforms like discrete cosine transform .In frequency domain, coefficients are slightly modified. This makes some unnoticeable changes in the whole image and makes it robust to attack [3,5].

2 Existing Algorithm for Watermarking in Spectral Domain

1. Calculate the 2-D Discrete cosine Transform of the asset to be watermarked with the message.[6][7].
2. Find out the M largest coefficients m_1, m_2, m_3 ……m_M from the above calculated discrete cosine transform .
3. Create a message(watermark) by generating a M-element pseudo-random sequence of numbers from Gaussian distribution μ=1 and variance=1 [2] [3] .
4. The message signal is been embedded into the M-largest discrete cosine transform coefficients.

Linear valued spectral processed formula (LVSP).

$$D_i = C_i \,.\, (1 + \alpha \,.w_i)$$

5. Calculate the inverse discrete cosine transformation of the result obtained in the above point.

2.1 Advanced Techniques for Watermarking in Spectral Domain

The main theme of this paper is modifying the equation in Step4 of existing algorithm by inducing more amount of message signal power in the asset and at the same time, keeping the quality of the image intact. And thereafter comparing the differences in the usage of discrete fourier transform over discrete cosine transform and their individual advantages are visualized through results.

2.2 Modified Algorithm for Watermarking in Spectral Domain

The modification is required in Step1 and Step4 of the existing algorithm.

Modified Step1: It is suggested to use both the discrete fourier transform and discrete cosine transform according to how the application demands and details shown pictorially below.

Modified Step2: The modified formula below will help in strengthening the message signal to be embedded in the asset (message to asset signal power ratio increases) and by modifying the values of the coefficients, we can directly increase the strength without going for higher strength signal.

n-valued spectral processed formula (NVSP) :

$$D_i = C_i \,.\, (1 + \beta_1.w_i + \beta_2 \,.w_i^2 + \beta_3.\, w_i^3 \; +\ldots\ldots\ldots + \beta_n \,.\, w_i^n) \qquad (1)$$

Extended n-valued spectral processed formula(ENVSP) :

$$D_i = C_i \,.\, (1 + \beta_1.w_i^2 + \beta_2 \,.\, w_i^4 + \beta_3 \,.\, w_i^6 \; +\ldots\ldots\ldots + \beta_n \,.w_i^{2n}) \qquad (2)$$

In Eqn1 and Eqn2, by including higher powers of watermark (w_i) ,the strength of watermark is increased in the resulting image obtained by NVSP and ENVSP method over the resulting image obtained by existing LVSP method by maintaining the same quality shown in the image results below.

The above equation ENVSP is favorable for $\alpha < 0.3$

For Eqn 1, if $\beta_1 = \alpha$, $\beta_2 = \alpha^2$, then

$$D_i = C_i \,.\, (1 + \alpha \,.w_i + \alpha^2 \,.w_i^2 + \alpha^3 .w_i^3 \ldots\ldots\ldots\ldots + \alpha^n \,.w_i^n) \tag{3}$$

$$D_i = C_i \,.\, (\alpha^n \,.w_i^n - 1 \,) / (\, \alpha \,.w_i - 1) \tag{4}$$

For Eqn 2 ,if $\beta_1 = \alpha^2$, $\beta_2 = \alpha^4$, then

$$D_i = C_i \,.\, (\, 1 + \alpha^2 .w_i^2 + \alpha^4 .w_i^4 \ldots\ldots\ldots\ldots + \alpha^{2n} .w_i^{2n}) \tag{5}$$

$$D_i = C_i \,.\, (\alpha^{2n} \,.w_i^{2n} - 1 \,) / (\alpha^2 \,.w_i^2 - 1) \tag{6}$$

So, if n value is been taken as infinite, then that is the case in maximum amount of watermark signal is been induced in the image but the problem in this case is it produces low quality images for higher values of α. But for lower values of α, the equations below clearly depict that this method is better improvement over the existing LVSP because in this case , D_i contains maximum number of higher powers of w_i as n is equal to infinite thereby inducing higher amount of watermark into the image.

Infinite valued spectral processed formula (IVSP) :

$$D_i = C_i \,.\, (\, 1 \,) / (1 - \alpha \,.w_i) \tag{7}$$

Extended Infinite valued spectral processed formula (EIVSP) :

$$D_i = C_i \,.\, (\, 1 \,) / (1 - \alpha^2 .w_i^2) \tag{8}$$

Generally, value of α is taken to be 0.1 and value of M = 1000.

Practical results show IVSP is efficient over previous existing LVSP for $\alpha < 0.4$ and for those cases above $\alpha > 0.4$, NVSP is used to get better results over LVSP. Practical results show IVSP is efficient over previous existing LVSP for $\alpha < 0.4$ and for those cases above $\alpha > 0.4$, NVSP is used to get better results over LVSP.

2.3 Resistance of IVSP, NVSP Methods in Overcoming the Factors Affecting the Digital Watermark Strength

- Image variations: For efficient embedding of a digital watermark into the asset, the factors on which it is dependent is variations and randomness. For this, we have to choose a higher digital watermark strength which can be clearly obtained by the method of IVSP and NVSP and their extended versions.
- Image size: The larger the size of image, the greater the number of pixels in the image, the more the digital watermark can be repeated through it. Since in the proposed advanced technique, we are only increasing the strength of each pixel of watermark, so the size hardly has any effect on the resulting image by this method.
- No effect of attacks: Robust digital image watermarking method proves to overcome geometrical attacks. Generally, since the strength of watermark signal produced by IVSP and NVSP is very high, so there is a minimal chance of attacks . But if minimal probability case occurs, then robust digital image watermarking method proves to be most effective solution additional to this.

Fig. 1. (a) Original Image A (b) Applied LVSP on A, α = 0.1 (c) Applied LVSP on A, α=0.2 (d) Applied LVSP on A, α =0.3 (e) Applied LVSP on A, α=0.5 (f) Applied IVSP on A, α =0.1 (g) Applied IVSP on A, α =0.2 (h) Applied IVSP on A, α =0.3 (h) Applied IVSP on A, α =0.4

2.4 Advantages of IVSP, LVSP and Their Extended Versions over Existing LVSP

The image results above in Fig.1 clearly depict the above advanced technique. The image obtained by implementing IVSP with α=0.1 is almost same quality as original image A than that of LVSP with α=0.1 when compared with the original image A and having highly embedded message signal because as the equations above clearly depict that as we are taking higher powers of w_i , so the overall value of D_i contains higher strength of watermark in the image and at the same time ,maintaining and retaining the quality of image. By quality of image, we mean that the traceability of the watermark by naked eye is not possible and the watermarked image should be similar to that of the asset used. As expected IVSP for α >0.3 is a poor quality watermarked image. So in that case, NVSP with low or moderate n value is preferable and thereby strengthening the signal power.

So, by the strength of the signal power, we mean that the message signal is induced much into the asset but the comparison is not always by magnitude. In IVSP, more amount of message signal power is induced into asset than in LVSP. But, it cannot

always be true in terms of magnitude. That is the main misconception we need to avoid because the message (watermark) signal which we are using is a mixture of pseudo numbers which can be positive or negative. So, if we specifically want even in terms of magnitude, then we can construct the image by using the modified formula EIVSP and increase the magnitude of signal and if suppose the quality of image goes down, then we can opt ENVSP so that we can maintain the quality of image

2.5 Comparisons between DCT and DFT in Watermarking Applied on Spectral Domain of Images

Both the transforms are highly preferable. The retrieved images in both cases appears to resemble each other but the image obtained by the taking the difference of the image obtained by DCT and DFT gives valuable information. As we know, in DCT, the coefficients are highly concentrated at a single area .Unlike DCT, in DFT, the coefficients are uniformly concentrated but the difference shows something unusual.

Fig. 2. (a) Applied LVSP on A, $\alpha = 0.1$, using Fourier Transform (b) Applied IVSP on A, $\alpha = 0.1$, using Fourier Transform

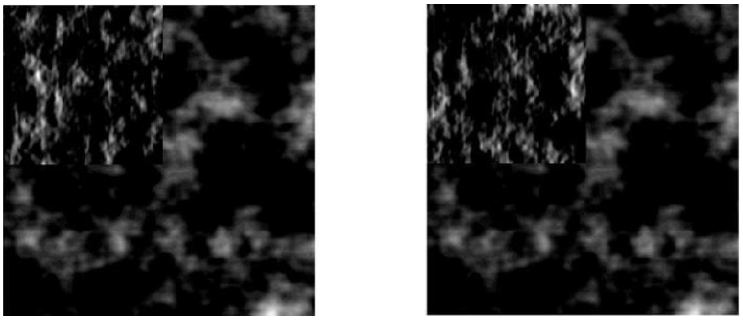


Fig. 3. (a) Image is obtained by taking the difference of two images : Image obtained on applying LVSP, $\alpha = 0.1$, using Discrete Fourier transform and Image obtained on applying IVSP $\alpha = 0.1$,using Discrete Cosine Transform .(b) Image is obtained by taking the difference of two images: Image obtained on applying IVSP, $\alpha = 0.1$, using Discrete Fourier transform and Image obtained on applying IVSP $\alpha = 0.1$,using Discrete Cosine Transform.

3 Conclusion

The above results thereby prove that the above proposed advanced technique is an effective technique over the existing technique and comparison of usage of different transforms enable the user to decide timely basis the kind of transform to be used for a particular application. The method can become even more elegant if we are able to design an algorithm for calculating the range of values of n in the proposed technique so that it can be generalized without any approximations. But IVSP is very well executed technique over LVSP and at the same time, IVSP induces high message signal power in asset.

References

1. Kim, J., Won, S., Zeng, W., Soohong: Copyright protection of vector map using digital watermarking in spatial domain content. In: 2011 7th International conference on Multimedia and its Applications (December 2011)
2. Syed, E.: Final Report of Digital Watermarking. University of Texas at Arlington (2011)
3. Ingemar, J.: Cox: Digital watermarking and steganography. Morgan Kaufmann, Burlington (2008)
4. Frank, Y.: Shih: Digital watermarking and steganography: fundamentals and techniques. Taylor & Francis, Boca Raton (2008)
5. Bendens, O.: Geometry-based Watermarking of 3D Models. IEEE Computer Graphics and Applications 19, 46–55 (1999)
6. Brown, L.G.: A Survey of Image Registration Techniques. ACM Computing Surveys, 325–376 (1992)
7. Bracewell, R.N.: The Fourier Transformations and its Applications. McGrawHill, NewYork (1986)

Materialized View Selection Using Memetic Algorithm

T.V. Vijay Kumar and Santosh Kumar

School of Computer and Systems Sciences,
Jawaharlal Nehru University,
New Delhi-110067, India

Abstract. A data warehouse stores historical data for the purpose of answering strategic and decision making queries. Such queries are usually exploratory and complex in nature and have high response time when processed against a continuously growing data warehouse. These response times can be reduced by materializing views in a data warehouse. These views, which contain pre-computed and summarized information, aim to provide answers to decision making queries in an efficient manner. All views cannot be materialized due to space constraints. Also, optimal view selection is shown to be an NP-Complete problem. Alternatively, several view selection algorithms exist, most of these being empirical or based on heuristics like greedy, evolutionary etc. In this paper, a memetic view selection algorithm, that selects the Top-T views from a multi-dimensional lattice, is proposed. This algorithm incorporates the local search improvement heuristic, i.e. Iterative Improvement, into the evolutionary manner for selecting an optimal set of views, from amongst all possible views, in a multidimensional lattice. The purpose is to efficiently select good quality views. This algorithm, in comparison to the better known greedy view selection algorithm, is able to efficiently select better quality views for higher dimensional data sets.

Keywords: Data Warehouse, Materialized View Selection, Memetic Algorithm.

1 Introduction

Voluminous data is available in data sources spread across the globe. Organizations, in order to be competitive, are continuously evolving their strategies for accessing and exploiting this data in an effective and efficient manner. Several approaches exist for accessing this data from the underlying data sources. These are mainly categorized into two types namely the Lazy, or on-demand approach, and the Eager, or in-advance approach [45]. In the former, relevant data, for processing the query, is extracted from the data sources whereupon the query is processed against the same. This delays the query processing, on account of the time consumed in the extraction of data, from the data sources, and processing the query against it. Whereas, in the latter approach, data is extracted and stored aprior in a central repository and any future query is processed against this central repository. Data warehousing is based on the latter approach and the central repository that stores data is referred to as the data warehouse[45]. A data warehouse stores data that is subject oriented, integrated, time variant and non-volatile and is created for the purpose of supporting decision making[17]. A data

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 316–327, 2013.
© Springer International Publishing Switzerland 2013

warehouse contains historical data accumulated over a period of time. This data, reflecting the past information querying trends, can be useful in devising strategies for efficient decision making. Decision making queries are usually ad-hoc, analytical and exploratory in nature and their response times are high when processed against the continuously growing data warehouse. This leads to delay in decision making. Materialized views [29] have been used as an alternative to address this problem.

Materialized views, unlike virtual views, store pre-computed aggregated and summarized information with the aim of providing answers to analytical queries in comparatively reduced response times. This would necessitate that the materialized views contain the relevant and required information for answering analytical queries and that these views should fit within the available space for materialization i.e. should conform to the space constraint [5]. All possible views cannot be materialized, as the number of views are exponential with respect to the number of dimensions. It thus would not be able to conform to the space constraints[14]. Further, selection of an optimal subset of views is shown to be an NP-Complete problem [14]. Thus, the only alternative available is to select a subset of views, from amongst all possible views that improves the query response time and fits within the available space for materialization. Selecting such a subset of views is referred to as the view selection problem [5]. View selection is concerned with the identification and selection of beneficial subsets of views, from amongst all possible views, in order to reduce the response time of analytical queries, even while conforming to resource constraints like storage space, memory and CPU usage etc [5, 11, 46, 47]. Several view selection approaches exist in literature, most of which are empirical based [1, 3, 4, 9, 20, 21, 22, 31, 32, 37]; or based on heuristics like greedy [11, 12, 13, 14, 30, 33, 34, 35, 36, 38, 39, 40, 41], evolutionary[16, 19, 43, 48, 49] etc. Majority of the view selection algorithms are greedy based and are focused around the greedy algorithm given in [14], which hereafter in this paper would be referred to as HRUA. HRUA selects the *Top-T* views from a multidimensional lattice, arrived at from a star schema representation of data in a data warehouse. Greedy algorithms are not scalable, as they are unable to select views for higher dimensional data sets. Alternatively, views can be selected in an evolutionary manner using the memetic algorithm (MA) [24]. MA adds the local search improvement heuristic into the evolutionary nature of the algorithm with the purpose of efficiently generating good quality solutions. The local improvement heuristic is applied to individuals in the population before exploring and exploiting the search space in an evolutionary manner. This improvement heuristic enables individuals in the population to gain experience before getting involved in the evolutionary process[8, 24, 50]. MA, which has been widely and successfully applied to solve combinatorial optimization problems, has an advantage in terms of its ease of implementation, intensive power of a local search and computational efficiency over other evolutionary algorithms[50]. An attempt has been made in this paper to use MA to address the materialized view selection problem.

In this paper, a memetic view selection algorithm (MVSA), that selects the *Top-T* views, from amongst all possible views in a multi-dimensional lattice, is proposed. MVSA selects views using MA by applying a local search improvement heuristic while selecting views in an evolutionary manner. MVSA is compared with HRUA, on the total cost of evaluating all the views (*TVEC*) selected by the two algorithms. MVSA is able to select comparatively better quality views.

The paper is organized as follows: The proposed view selection algorithm MVSA along with an example is given in section 2. Experimental results are given in section 3. Section 4 is the conclusion.

2 View Selection Using Memetic Algorithm

As discussed above, it is infeasible to select all possible views due to space constraints. Also an optimal selection of views, from amongst all possible views, is an NP Complete problem. Thus, there is a need to select a good set of views, from amongst all possible views, in a multidimensional lattice. In this paper, the memetic algorithm has been used to select the *Top-T* views from a multi-dimensional lattice[14]. The memetic algorithm is discussed next.

2.1 Memetic Algorithm

According to [6, 26], human behavior can be decomposed into memes, which are simple units of imitation in cultural transmission. A meme is a unit of knowledge added to other memes for generating a new meme, which is likely to be more interesting and can be easily propagated within the human community. In humans, some memes may not be important and useful, and these gradually die out. The ability of memes to modify or evolve themselves during their life time makes them different from genes. The lifetime learning of a meme enables it to adapt faster than a gene. This interesting aspect of meme has been a major inspiration behind the memetic algorithm[8, 26]. The memetic algorithm belongs to a class of stochastic global search techniques that combine, within the framework of Evolutionary Algorithms, the benefits of problem-specific local search heuristics. Memetic algorithms are based on populations of individuals representing the candidate solutions. These are composed of an evolutionary approach, with a set of local search algorithms used within each cycle of the genetic algorithm (GA)[15, 23]. It extends GA by applying a local search improvement heuristic to individuals of a population in each generation. It has been successfully used to solve a wide range of combinatorial optimization problems such as discrete, continuous, constrained, multi-objective etc.[26]. Hill climbing is a widely used local search improvement heuristic for solving these problems [2, 27, 28]. A general memetic algorithm [8] is given in Fig. 1.

Step0: Define a genetic representation of an individual for the given optimization problem
Step1: Create N individuals of the population P
Step2: Each individual undergoes improvement using local improvement heuristic
Step3: Select individuals, based on their fitness, for crossover and mutation
Step4: Perform crossover and mutation
Step5: Replace individuals in P by new individuals
Step6: If termination condition not reached then GO TO Step 2

Fig. 1. Memetic Algorithm[8]

In the memetic algorithm given in Fig. 1, first a solution representation of an individual in the population is defined for the given optimization problem. It should reflect the problem and its fitness can be computed from it. Based on this representation, individuals in the population are generated. Thereafter, each individual solution is improved using some local search improvement heuristic. This is followed by selecting individuals in the population, based on their fitness, for crossover and mutation. After crossover and mutation are performed, a population with a new set of individuals is generated. This process (Step2 to Step5) is repeated until the termination condition is reached. The termination condition may be a pre-specified number of generations or an acceptable solution has been achieved, or there is no improvement in the solution for a pre-specified number of generations.

In this paper, algorithm MVSA has been proposed that selects the *Top-T* views, from amongst all possible views in a multidimensional lattice, using the memetic algorithm. MVSA is discussed next.

2.2 MVSA

The proposed algorithm MVSA that selects the *Top-T* views, from amongst all possible views in a multidimensional lattice, is given in Fig. 2.

Input: Lattice *L* of views with size of each view, Number of Generations *G*,
Set of views to materialize *Top-T*, Initial population size *PopSize*
Output: *Top-T* views
Method:
1. //Generate initial population *Pop* with size *PopSize* with chromosome representation for *Top-T* views from Lattice *L* as
$\{V_1, V_2, V_3, \ldots, V_{\text{Top-T}}\}$
2. // Improve the *Top-T* views in *Pop* using local-search improvement heuristic
FOR i=1 to *PopSize*
IPop[i] = ImprovementHeuristic(*Pop*[i])
END FOR
3. WHILE Generation< G
DO
(i) Compute TVEC of each *Top-T* views in *IPop* using the following formula

$$TVEC = \sum_{i=1 \wedge SM_{V_i}=1}^{N} Size(V_i) + \sum_{i=1 \wedge SM_{V_i}=0}^{N} SizeSMA(V_i)$$

where N is total number of Views in the Lattice, Size(V_i) is size of view V_i,
SM_{V_i} is the StatusMaterialized of view V_i ($SM_{V_i} = 1$, if materialized, $SM_{V_i} = 0$, if not materialized)
SizeSMA(Vi) is size of smallest materialized ancestor of view Vi
(ii) Select a set of *Top-T* views from *IPop* using binary tournament selection
(iii) Perform cyclic crossover with probability P_c and mutation with probability P_m on selected *Top-T* views
(iv) Assign the new population of *Top-T* views to *Pop*
(v) FOR i=1 to *PopSize*
IPop[i] = ImprovementHeuristic(*Pop*[i])
END FOR
(vi) Increment Generation by 1
END DO
4. RETURN *Top-T* Views

Fig. 2. Algorithm MVSA

MVSA considers the lattice *L* of views, with the size of each view, number of generations the algorithm is to run *G*, the set of views to materialize *Top-T* and the population size *PopSize*. It produces the *Top-T* views having the minimum TVEC as output. First the initial population *Pop*, of size *PopSize*, of the *Top-T* views is generated randomly from lattice *L*. Next, a local search improvement heuristic is applied to each of the *Top-T* views in the population in order to improve the initial set of the *Top-T* views. In MVSA, Iterative Improvement [18, 42] is applied as the local search improvement heuristic and is given in Fig. 3. Several other local search improvement heuristics like hill climbing[2, 27, 28], simulated annealing [18, 25, 44], two-phase optimization [18] etc. can also be applied.

```
BEGIN
        SET an initial value of Top-T views minV_T with very high TVEC
        WHILE not (stop_condition) DO
            Select a random Top-T views V_T
            WHILE r-local minimum not reached DO
                V_T' = Neighbor(V_T)
                IF TVEC(V_T') < TVEC (V_T) THEN V_T ← V_T'
            END DO
            IF TVEC (V_T) < TVEC (minV_T) then minV_T = V_T
        END DO
        RETURN (minV_T)
END
```

Fig. 3. Iterative Improvement Algorithm[18, 42]

The *TVEC* of the improved candidate set of *Top-T* views is then computed using the following formula[42, 43, 44]:

$$TVEC = \sum_{i=1 \wedge SM_{V_i}=1}^{N} Size(V_i) + \sum_{i=1 \wedge SM_{V_i}=0}^{N} SizeSMA(V_i)$$

where

N is total number of Views in the Lattice

Size(Vi) is size of view V_i

SizeSMA(Vi) is size of smallest materialized ancestor of view *Vi*

SM_{V_i} is the Status Materialized of view V_i ($SM_{V_i} = 1$, if materialized, $SM_{V_i} = 0$, if not materialized)

Next, the *Top-T* views are selected for crossover and mutation from the population *Pop* using the binary tournament selection[10] given in Fig. 4.

```
Initialize a parameter k with a value between 0 and 1
Choose two Top-K views randomly from the population
Choose a random number r between 0 and 1.
If r < k
   Select the Top-K views with lower TVEC
Else
   Select the Top-K views with higher TVEC
```

Fig. 4. Binary Tournament Selection Method[10]

Crossover for permutation[7], with probability P_c, and mutation[7], with probability P_m, are performed on the selected *Top-T* views. The crossover and mutation operations are performed as shown in Fig. 5.

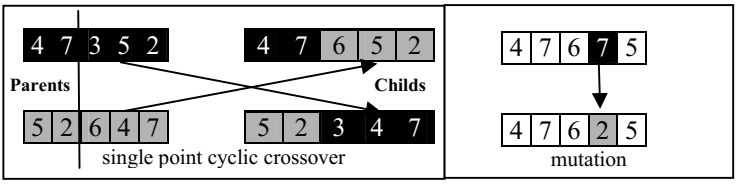


Fig. 5. Crossover and Mutation

The local search improvement heuristic, Iterative Improvement, is then applied to improve the set of *Top-T* views produced after crossover and mutation. This process of selection, crossover and mutation followed by an improvement of the *Top-T* views in the population *Pop* using the Iterative Improvement continues for a pre-specified number of generations *G*. Thereafter, the *Top-T* views, having the minimum *TVEC*, are produced as output.

Next, an example is given that illustrates the use of MVSA for selecting the *Top-T* views for materialization.

2.3 An Example

Consider the selection of *Top-4* views from a 3-dimensional lattice shown in Fig. 6.

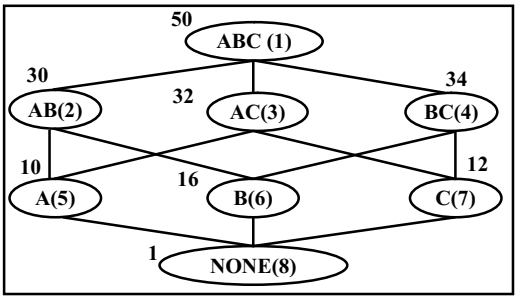


Fig. 6. 3-dimensional lattice along with size and index of each view

From this lattice, first an initial population *Pop* of *Top-4* views is randomly generated and is given in Fig. 7.

***Top-4* Views**	**Chromosomes (*Top-4* Views)**
AC, AB, B, C	3 2 6 7
AB, BC, A, C	2 4 5 7
BC, A, B, C	4 5 6 7
AB, AC, BC, A	2 3 6 5

Fig. 7. Initial population of *Top-4* views

Next, local search improvement heuristic Iterative Improvement is applied to each of the *Top-4* views in the population. For this, the *TVEC* of the set of *Top-4* views are computed. The *TVEC* computation of the *Top-4* views (3, 2, 6, 7) is shown in Fig. 8. These *Top-4* views are improved using Iterative Improvement, as shown in Fig. 9. In a similar manner, the *Top-4* views (2, 4, 5, 7), (4, 5, 6, 7) and (2, 3, 4, 5) are improved. These improved *Top-4* views, along with their *TVEC*, are given in Fig. 10.

$$\sum_{i=1 \wedge SM_{V_i}=1}^{8} Size(V_i) = (Size(ABC) + Size(AC) + Size(AB) + Size(B) + Size(C)) = (50 + 32 + 30 + 16 + 12) = 140$$

$$\sum_{i=1 \wedge SM_{V_i}=0}^{8} SizeSMA(V_i) = (SizeSMA(BC) + SizeSMA(A) + SizeSMA(None)) = (50 + 30 + 12) = 92$$

$$TVEC = \sum_{i=1 \wedge SM_{V_i}=1}^{8} Size(V_i) + \sum_{i=1 \wedge SM_{V_i}=0}^{8} SizeSMA(V_i) = 140 + 92 = 232$$

Fig. 8. TVEC computation of *Top-4* views (3, 2, 6, 7)

V_T	$TVEC(V_T)$	V_T'	$TVEC(V_T')$	$minV_T$	*Top-4* Views
3 2 6 7	232	3 2 6 5	230	224	AC, AB, B, C
3 2 6 5	230	3 2 7 5	224		
3 2 7 5	224	3 4 7 5	232		
3 2 7 5	224	3 2 4 5	228		
3 2 7 5	224	4 2 7 5	228		

Fig. 9. Iterative Improvement on *Top-4* views (3, 2, 6, 7)

Top-4 Views	*TVEC*	*Top-4* Views after II	*TVEC*
3 2 6 7	232	3 2 7 5	224
3 4 6 5	234	3 2 6 5	230
4 5 6 7	232	4 5 2 7	226
2 3 6 5	230	2 3 4 5	228

Fig. 10. Top-4 views in the population after Iterative Improvement (II)

Next, the *Top-4* views are selected for crossover and mutation using the binary tournament selection, as given in Fig. 11. The selected *Top-4* views undergo crossover with the probability P_c=0.75 and mutation with the probability P_m=0.1. These are shown in Fig. 12. The *Top-4* views after crossover and mutation are improved using Iterative Improvement as given in Fig. 13.

Tournament between individuals [P(i)] & [P(j)]	[*TVEC*(P(i)] & [*TVEC*(P(j)]	Random (r)	*Top-4* views Selected
3 2 7 5 & 2 3 4 5	224 & 228	0.44	3 2 7 5
3 2 6 5 & 4 5 2 7	230 & 226	0.88	3 2 6 5
4 5 2 7 & 2 3 4 5	226 & 228	0.66	4 5 2 7

Fig. 11. Selection of *Top-4* views using using Binary Tournament Selection

Top-4 views for Crossover	Crossover Point	*Top-4* views after Crossover
3 2 7 5	2	3 2 7 4
4 5 2 7		4 5 7 3
3 2 6 5	1	3 5 2 7
4 5 2 7		4 2 6 5

Top-4 views for Mutation	Mutation Point	*Top-4* views after Mutation
3 2 7 4	-	3 2 7 4
4 5 7 3	-	4 5 7 3
3 5 2 7	-	3 5 2 7
4 2 6 5	3	4 2 7 5

Fig. 12. Crossover and Mutation on *Top-4* views

Top-4 Views	*TVEC*	*Top-4* Views after II	*TVEC*
3 2 7 4	230	5 2 7 4	226
4 5 7 3	232	4 5 2 3	228
3 5 2 7	224	3 5 2 7	224
4 2 7 5	226	3 2 7 5	224

Fig. 13. Iterative Improvement (II) on *Top-4* views in the population

The above process is repeated for a pre-specified number of generations, whereafter the *Top-4* views having minimum *TVEC* is produced as output.

3 Experimental Results

Algorithms MVSA and HRUA were implemented using JDK 1.6 in Windows-7 environment. The two algorithms were compared by conducting experiments on an Intel based 2.13 GHz PC having 3 GB RAM. The comparisons were carried out on the *TVEC* due to views selected by the two algorithms.

First, graphs showing the *TVEC*, for different crossover and mutation probabilities for selecting the *Top-10* views after 100 generations, were plotted and compared with those selected using HRUA. These graphs, for the pair of crossover (P_c) and mutation (P_m) probabilities (0.7, 0.05), (0.7, 0.1), (0.8, 0.05), (0.8, 0.1), are shown in Fig. 14.

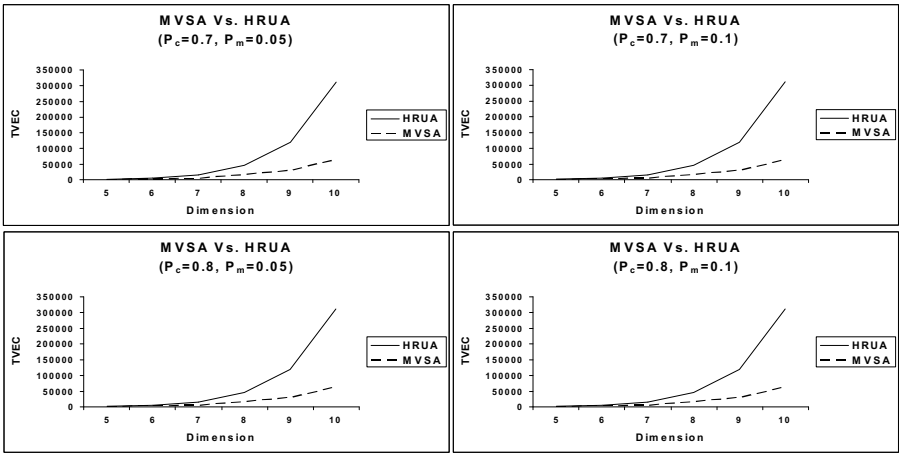


Fig. 14. MVSA Vs. HRUA – *TVEC* Vs. Dimensions for different P_c's and P_m's

The graphs show that MVSA, in comparison to HRUA, is able to select views at a lower *TVEC* for different crossover and mutation probabilities and this difference becomes maximum across all dimensions for P_c=0.8 and P_m=0.05. Accordingly, for these observed crossover and mutation probabilities P_c=0.8 and P_m=0.05, graphs showing *TVEC* of the *Top-T* views selected after 100 generations for dimensions 7, 8, 9, 10, were plotted. These graphs are shown in Fig. 15.

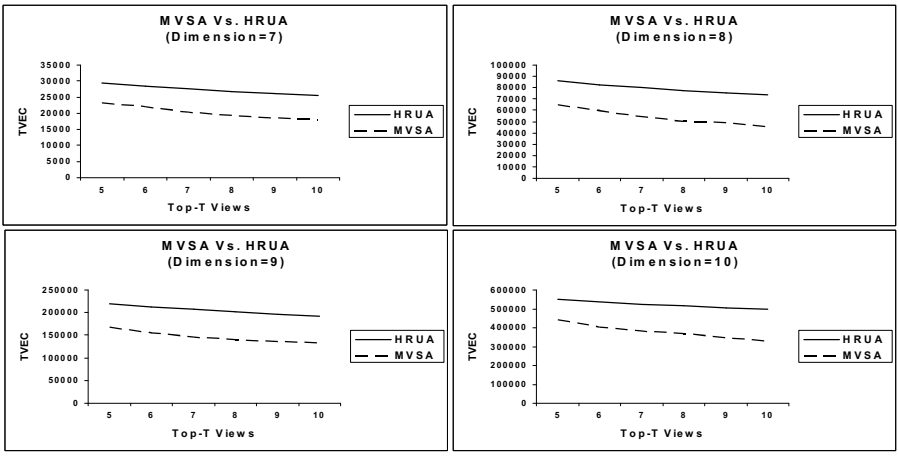


Fig. 15. MVSA Vs. HRUA – *TVEC* Vs. *Top-T* Views for P_c=0.8 and P_m=0.05

It can be noted from the graph that the *TVEC* of the *Top-T* views, for each value of *T*, selected using MVSA is lesser than those selected using HRUA. This difference becomes significant for higher values of *T*. Thus it can be inferred that MVSA performs better than HRUA with respect to the quality of views selected for materialization for the observed crossover and mutation probability of 0.8 and 0.05 respectively.

4 Conclusion

In this paper, an algorithm MVSA for selecting the *Top-T* views, from amongst all possible views, in a multidimensional lattice is presented. MVSA, which selects views using the memetic algorithm, adds the local search improvement heuristic Iterative Improvement into the evolutionary nature of the algorithm in order to efficiently select views having a lower *TVEC*. MVSA, in each iteration, applies Iterative Improvement to the *Top-T* views in the population before exploring and exploiting the search space in an evolutionary manner. This could lead to an efficient selection of good quality views, i.e. views having lower *TVEC*. Further experimental results show that MVSA, in comparison to the well known greedy algorithm HRUA, selects the *Top-T* views at a comparatively lower *TVEC* for the observed crossover and mutation probabilities. That is, MVSA is able to select comparatively better quality views, which, when materialized, would reduce the query response time and lead to efficient decision making.

References

1. Agrawal, S., Chaudhari, S., Narasayya, V.: Automated Selection of Materialized Views and Indexes in SQL databases. In: 26th International Conference on Very Large Data Bases (VLDB 2000), Cairo, Egypt, pp. 495–505 (2000)
2. Alkan, A., Ozcan, E.: Memetic Algorithms for Timetabling, IEEE Congress on Evolutionary Computation, pp. 1796–1802 (2003)
3. Aouiche, K., Darmont, J.: Data mining-based materialized view and index selection in data warehouse. Journal of Intelligent Information Systems, 65–93 (2009)
4. Baralis, E., Paraboschi, S., Teniente, E.: Materialized View Selection in a Multidimansional Database. In: 23rd International Conference on Very Large Data Bases (VLDB 1997), Athens, Greece, pp. 156–165 (1997)
5. Chirkova, R., Halevy, A.Y., Suciu, D.: A Formal Perspective on the View Selection Problem. Proceedings of VLDB, 59–68 (2001)
6. Dawkins, R.: The Selfish Gene. Clarendon Press, Oxford (1976)
7. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. Springer (2003)
8. Elbeltagi, E., Hegazy, T., Grierson, D.: Comparison among five evolutionary-based optimization algorithms. Advanced Engineering Informatics, 19, 43–53 (2005)
9. Golfarelli, M., Rizzi, S.: View Materialization for Nested GPSJ Queries. In: Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW 2000), Stockholm, Sweden (2000)
10. Goldberg, D.E., Deb, K.: A comparative analysis of selection schemes used in Genetic Algorithms. Foundations of Genetic Algorithms, MK, 69–93 (1991)
11. Gupta, H., Mumick, I.S.: Selection of Views to Materialize in a Data warehouse. IEEE Transactions on Knowledge & Data Engineering 17(1), 24–43 (2005)
12. Gupta, H., Harinarayan, V., Rajaraman, V., Ullman, J.: Index Selection for OLAP. In: Proceedings of the 13th International Conference on Data Engineering, ICDE 1997, Birmingham, UK (1997)
13. Haider, M., Vijay Kumar, T.V.: Materialised Views Selection using Size and Query Frequency. International Journal of Value Chain Management (IJVCM) 5(2), 95–105 (2011)
14. Harinarayan, V., Rajaraman, A., Ullman, J.D.: Implementing Data Cubes Efficiently. In: ACM SIGMOD, Montreal, Canada, pp. 205–216 (1996)
15. Hart, W.E., Krasnogor, N., Smith, J.E.: Memetic evolutionary algorithms. In: Hart, W.E., Krasnogor, N., Smith, J.E. (eds.) Recent Advances in Memetic Algorithms, pp. 3–27. Springer, Berlin (2004)
16. Horng, J.T., Chang, Y.J., Liu, B.J., Kao, C.Y.: Materialized View Selection Using Genetic Algorithms in a Data warehouse System. In: Proceedings of the 1999 congress on Evolutionary Computation, Washington, D. C., USA, vol. 3 (1999)
17. Inmon, W.H.: Building the Data Warehouse, 3rd edn. Wiley Dreamtech India Pvt. Ltd (2003)
18. Ioannidis, Y.E., Kang, Y.C.: Randomized Algorithms for Optimizing Large Join Queries. In: Proceedings of the 1990 ACM Sigmod International Conference on Management of Data, vol. 19(2), pp. 312–321. ACM SIGMOD Record (1990)
19. Lawrence, M.: Multiobjective Genetic Algorithms for Materialized View Selection in OLAP Data Warehouses. In: GECCO 2006, Seattle Washington, USA, July 8-12 (2006)
20. Lehner, W., Ruf, T., Teschke, M.: Improving Query Response Time in Scientific Databases Using Data Aggregation. In: Thoma, H., Wagner, R.R. (eds.) DEXA 1996. LNCS, vol. 1134, Springer, Heidelberg (1996)

21. Lin, Z., Yang, D., Song, G., Wang, T.: User-oriented Materialized View Selection. In: The 7th IEEE International Conference on Computer and Information Technology (2007)
22. Luo, G.: Partial Materialized Views. In: International Conference on Data Engineering (ICDE 2007), Istanbul, Turkey (April 2007)
23. Mitchell, M.: An Introduction to Genetic Algorithms. The MIT Press (1999)
24. Moscato, P.: On evolution, search, optimization, genetic algorithms and martial arts: towards memetic algorithms, Technical Report Caltech Concurrent Computation Program. California Institute of Technology, Pasadena (1989)
25. Nahar, S., Sahni, S., Shragowitz, E.: Simulated Annealing and Combinatorial Optimization. In: Proceedings of 23rd Design Automation Conference, pp. 293–299 (1986)
26. Neri, F., Cotta, C.: Memetic algorithms and memetic computing optimization: A literature review. Swarm and Evolutionary Computation 2, 1–14 (2012)
27. Ozcan, E., Mohan, C.K.: Steady State Memetic Algorithm for Partial Shape Matching. In: 7th Annual Conference on Evolutionary Programming, pp. 527–536 (1998)
28. Ozcan, E., Onbasioglu, E.: Genetic Algorithms for Parallel Code Optimization. In: IEEE Congress on Evolutionary Computation (2004)
29. Roussopoulos, N.: Materialized Views and Data Warehouse. In: 4th Workshop KRDB 1997, Athens, Greece (August 1997)
30. Shah, B., Ramachandran, K., Raghavan, V.: A Hybrid Approach for Data Warehouse View Selection. International Journal of Data Warehousing and Mining 2(2), 1–37 (2006)
31. Teschke, M., Ulbrich, A.: Using Materialized Views to Speed Up Data Warehousing, Technical Report, IMMD 6, Universität Erlangen-Nümberg (1997)
32. Theodoratos, D., Sellis, T.: Data Warehouse Configuration. In: Proceeding of VLDB, Athens, Greece, pp. 126–135 (1997)
33. Valluri, S., Vadapalli, S., Karlapalem, K.: View Relevance Driven Materrialized View Selection in Data Warehousing Environment. Australian Computer Science Communications 24(2), 187–196 (2002)
34. Vijay Kumar, T.V., Ghoshal, A.: A reduced lattice greedy algorithm for selecting materialized views. In: Prasad, S.K., Routray, S., Khurana, R., Sahni, S. (eds.) ICISTM 2009. Communications in Computer and Information Science, vol. 31, pp. 6–18. Springer, Heidelberg (2009)
35. Vijay Kumar, T.V., Haider, M., Kumar, S.: Proposing candidate views for materialization. In: Prasad, S.K., Vin, H.M., Sahni, S., Jaiswal, M.P., Thipakorn, B. (eds.) ICISTM 2010. Communications in Computer and Information Science, vol. 54, pp. 89–98. Springer, Heidelberg (2010)
36. Kumar, T.V.V., Haider, M.: A Query Answering Greedy Algorithm for Selecting Materialized Views. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS, vol. 6422, pp. 153–162. Springer, Heidelberg (2010)
37. Vijay Kumar, T.V., Goel, A., Jain, N.: Mining Information for Constructing Materialised Views. International Journal of Information and Communication Technology 2(4), 386–405 (2010)
38. Vijay Kumar, T.V., Haider, M.: Greedy views selection using size and query frequency. In: Unnikrishnan, S., Surve, S., Bhoir, D. (eds.) ICAC3 2011. CCIS, vol. 125, pp. 11–17. Springer, Heidelberg (2011)
39. Vijay Kumar, T.V., Haider, M., Kumar, S.: A view recommendation greedy algorithm for materialized views selection. In: Dua, S., Sahni, S., Goyal, D.P. (eds.) ICISTM 2011. CCIS, vol. 141, pp. 61–70. Springer, Heidelberg (2011)

40. Vijay Kumar, T.V., Haider, M.: Selection of views for materialization using size and query frequency. In: Das, V.V., Thomas, G., Lumban Gaol, F. (eds.) AIM 2011. CCIS, vol. 147, pp. 150–155. Springer, Heidelberg (2011)
41. Vijay Kumar, T.V., Haider, M.: Materialized Views Selection for Answering Queries. In: Kannan, R., Andres, F. (eds.) ICDEM 2010. LNCS, vol. 6411, pp. 44–51. Springer, Heidelberg (2012)
42. Vijay Kumar, T.V., Kumar, S.: Materialized view selection using iterative improvement. In: Meghanathan, N., Nagamalai, D., Chaki, N. (eds.) Advances in Computing & Inf. Technology. AISC, vol. 178, pp. 205–213. Springer, Heidelberg (2012)
43. Vijay Kumar, T.V., Kumar, S.: Materialized view selection using genetic algorithm. In: Parashar, M., Kaushik, D., Rana, O.F., Samtaney, R., Yang, Y., Zomaya, A. (eds.) IC3 2012. CCIS, vol. 306, pp. 225–237. Springer, Heidelberg (2012)
44. Vijay Kumar, T.V., Kumar, S.: Materialized View Selection Using Simulated Annealing. In: Srinivasa, S., Bhatnagar, V. (eds.) BDA 2012. LNCS, vol. 7678, pp. 168–179. Springer, Heidelberg (2012)
45. Widom, J.: Research Problems in Data Warehousing. In: 4th International Conference on Information and Knowledge Management, Baltimore, Maryland, pp. 25–30 (1995)
46. Yang, J., Karlapalem, K., Li, Q.: Algorithms for Materialized View Design in Data Warehousing Environment. The Very Large databases (VLDB) Journal, 136–145 (1997)
47. Yousri, N.A.R., Ahmed, K.M., El-Makky, N.M.: Algorithms for Selecting Materialized Views in a Data Warehouse. In: The Proceedings of International Conference on Computer Systems and Applications, AICCSA 2005, pp. 27–1 (2005)
48. Zhang, C., Yao, X., Yang, J.: Evolving Materialized Views in a Data Warehouse. In: IEEE CEC, pp. 823–829 (1999)
49. Zhang, C., Yao, X., Yang, J.: An Evolutionary Approach to Materialized Views Selection in a Data Warehouse Environment. IEEE Transactions on Systems, Man and Cybernatics, 282–294 (2001)
50. Zhang, Q., Sun, X., Wang, Z.: An Efficient MA-Based Materialized Views Selection Algorithm. In: IEEE Intl. Conf. on Control, Automation and Systems Engineering (2009)

Forgery Detection in Ballistic Motion Videos Using Motion Estimation and Modelling

Jithin Raj and Madhu S. Nair

Department of Computer Science, University of Kerala, Kariavattom,
Thiruvanathapuram-695581, Kerala, India
jithinraj90@gmail.com, madhu_s_nair2001@yahoo.com

Abstract. In this paper, we propose a method for detection of forgery in ballistic motion videos using motion estimation and modelling. Motion between consecutive frames is estimated using block matching algorithm without interpolation and is represented by a motion vector. State matrices are formed from the motion vector and a Markov process model is applied to it to get the transition probability matrix. By analysing the probability values in this matrix, the transition from one frame to the next is evaluated. For authentic videos, the transition probability matrices for all pair of subsequent frames show uniform characteristics whereas for fake videos, we can see difference in these characteristics. Thus transition probability matrix is used here as a feature vector for classifying a ballistic motion video as an authentic one or fake. The method is evaluated using various original and fake ballistic motion videos and yields good results in both static and moving camera videos.

Keywords: Ballistic Motion, Video Forgery Detection, Motion Estimation, Markov Modelling.

1 Introduction

Digital multimedia applications have become an inevitable part of our day to-day life because of the advancements in internet technology and availability of low-cost multimedia devices. Video sharing websites make video as the primary medium of entertainment in digital world. Availability of softwares for video editing allow easy tampering and manipulation of digital video data in numerous ways without leaving visible clues about the forgery. As a result, the authenticity of video content cannot be guaranteed and here comes the significance of digital video forensics.

One such area in which video editing is usually done is ballistic motion videos where forged video sequences are used to make realistic dynamic motion of projectiles. For instance, forgery is done in ballistic sports videos to create videos of amazing performances. Ballistic motion or projectile motion is the free falling motion of a body thrown in air with only acceleration due to gravity acting on it. Hence its velocity in horizontal direction will be constant and the gravitational acceleration acts in vertical direction which controls the motion and gives

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 328–339, 2013.
© Springer International Publishing Switzerland 2013

it a parabolic trajectory. Ballistic motion videos include videos of spectacular basketball shots, acrobatics, leaps, jumps etc. Some of these videos are authentic, but some are fake. The statistical methods for video forgery detection that uses interlaced and de-interlaced correlations [1], sensor noise patterns [2-4] and double-compression artifacts [5-7] cannot be applied here because these videos in the internet have already suffered a lot from several compression and recompression techniques, pre-processing etc. Here comes the need to find a method that does not depend on such statistical parameters of video.

One previous approach to solve this issue was by a fully geometric technique in which the trajectory of the moving object is tracked from the video and checked whether that path is consistent with the geometry and physics of a free-falling object [8]. The experimental analysis shows that their method gives good results. The problem with this method is that it is entirely dependent on the tracked positions of the moving object in the video and hence the efficiency of this method is determined by how effectively the object can be tracked. More than that, videos taken by a moving camera need special care as the object position depends on the camera movement also. In such situations, a camera calibration must be done which needs the aid of some other camera calibration software like Voodoo Camera Tracker [9].

These issues can be addressed if we are able to measure the motion of the object in the video rather than just tracking the geometrical position of the object. Relative motion between the frames is the key factor that can be used to find out any manipulation done in the ballistic motion videos because ballistic motion (projectile motion) is not random and the motion of the object from time to time is determined by a mathematical model. In our method, the motion between the successive frames is measured and modelled. Various algorithms have been put forward in literature for motion estimation and here we use block matching without interpolation which is a variant of basic block-matching sub-pixel motion estimation technique [10]. For modelling, we use Markov process model as described in [11] for steganalysis. For authentic videos, the motion will be consistent with the model whereas for fake videos, estimated motion shows a variation from the model. This method can handle videos taken by both static and dynamic camera without any assistance of other camera calibration software.

2 Proposed Method

In this section, we present the details of the proposed forgery detection technique based on motion estimation and modelling. Motion between two adjacent frames in the entire video sequence is calculated using block matching algorithm and the resulting motion vectors are modelled using Markov modelling. A deviation from the transition probability model is the proof of fake in the video.

2.1 Motion Estimation

Motion estimation is the first and fundamental step of various video processing applications such as standards conversion, frame-rate up-conversion, noise

reduction,image stabilisation, mosaicing and artifact removal. More importantly it is a critical component of video coding and transmission systems where motion estimation and compensation are used for obtaining efficient compression. In this work, we find a new application of motion estimation by using it for video forgery detection.

The basic principle behind motion estimation is that the pixel intensity patterns corresponding to foreground and background in a frame change its position to form corresponding objects in the next frame of the video, which is a time sequence of images. In sub-pixel motion estimation, we estimate the displacement of image structures from one frame to the next frame in the video. In this work, we use block matching algorithm without interpolation [10] for sub-pixel motion estimation.

Block Matching Algorithm. In block matching, the target frame whose motion is to be estimated is divided into small non-overlapping blocks called macro blocks. To determine the displacement of a particular macro block in target frame ft at time t, the previous reference frame fr at time $t - \Delta t$ is searched for the best matching block of the same size. Important parameters to be considered while implementing a block matching algorithm are matching criterion, search procedure, block size, spatial resolution of the displacement field and amplitude resolution of the displacement field.

Usually matching is done by minimizing an error criterion. The most commonly used error criteria are the mean square error (MSE) and the minimum absolute difference (MAD). For a BB macro block MB, MSE and MAD can be defined as,

$$MSE(\Delta x, \Delta y) = \frac{1}{B^2} \sum_{(i,j)\in MB} [(i + \Delta x, j + \Delta y, t - \Delta t) - I(i, j, t)]^2 \quad (1)$$

$$MSE(\Delta x, \Delta y) = \frac{1}{B^2} \sum_{(i,j)\in MB} |(i + \Delta x, j + \Delta y, t - \Delta t) - I(i, j, t)| \quad (2)$$

where $I(i, j, t)$ represents the intensity value at pixel position (i, j) of frame ft at time t and d represents the maximum allowable absolute displacement which is termed as the search parameter, where $-d \leq \Delta x, \Delta y \leq d$. Thus we get a displacement measure $[d_x(u, v)d_y(u, v)]$ for each block (u, v) where $[\Delta x \Delta y]$ minimizes the error criterion.

To obtain the motion vector of the target frame of size $M \times N$ with respect to the previous reference frame, displacement is calculated for each block in both horizontal and vertical directions. Then we can form the horizontal motion vector MV_x and vertical motion vector MV_y of size $m \times n$ as,

$$MV_x(u, v) = d_x^{(u,v)} \text{ and } MV_y(u, v) = d_y^{(u,v)} \text{ where,}$$
$$m = \lceil \tfrac{M}{B} \rceil \text{ and } n = \lceil \tfrac{N}{B} \rceil$$

These two can be combined to get the motion vector MV as,

$$MV = \sqrt{MV_x^2 + MV_y^2} \quad (3)$$

Thus, motion vector of a frame stipulates the magnitude and direction of the movement of a block from one frame to the next. As the maximum allowable absolute displacement (search parameter) is set as d, we can say that the range of motion magnitude will be from d to d. Another thing to be noted is that the size of the motion vector is determined by the block size. If we select very small block size, size of motion vector increases but the accuracy will be reduced because matching block is selected with respect to very small local context window. But if we increase the block size very much, the size of motion vector gets very much reduced. In such cases, the search and matching become almost global and the motion vector fails to represent the local motion, which is the actual area of interest.

The efficiency of block matching algorithm is determined by how we search through the reference frame blocks to find the best match. We can search over all possible positions determined by the search parameter which is an exhaustive search called full search. It is the most computationally expensive search in which we have to search through $(2d+1)^2$ blocks for finding the best match for each block. Several non-exhaustive sub-optimal searching techniques have been proposed to reduce the search time and to increase the efficiency of matching algorithm. They include three step search, four step search, diamond search, circular search, logarithmic search, cross search, orthogonal search, gradient descent search etc. Literature study shows that 2-D logarithmic cross search gives best matching in less time complexity with $(5 + 4log_2 d)$ searches for each block[12].

In conventional block matching algorithms that are designed for video coding applications, search for the best matching block is done in an enlarged (interpolated) reference search area in order to get good sub-pixel accuracy and better motion compensation frames. This is computationally expensive as the number of searches required is directly proportional to the interpolation factor and the cost is higher for more precise motion vectors. But for our work which is not related to video compression, the interpolation process is an overhead as we do not need motion compensation frames. Hence we use a fast motion estimation algorithm that achieves sub-pixel accuracy without interpolation [10]. The search method adopted is 2-D logarithmic cross search [12].

2.2 Motion Modelling

The motion vectors that are obtained from motion estimation stage are modelled using Markov process model. Markov chains and Markov processes are important classes of stochastic processes. A Markov chain is a discrete-time process for which the future behaviour, given the past and the present, only depends on the present and not on the past. A Markov process is the continuous-time version of a Markov chain. A Markov process is characterized by a set of states S and the transition probabilities $P(n|m)$ between the states. Here, $P(n|m)$ is the probability that the process is at state n in the next time point, given that it is at state m in the present time point. The matrix T with elements $P(n|m)$ is called the transition probability matrix of the Markov process.

In order to obtain the transition probability matrix by applying Markov process to the motion in the video frames, first we have to select the transition states. These states are defined as the difference in the displacement magnitude in the motion vector for two adjacent blocks. Adjacency can be in horizontal, vertical, diagonal and minor-diagonal directions. Hence we get four different state matrices for each motion vector and each of them can be defined as,

$$MV_h(i,j) = MV(i,j) - MV(i+1,j) \tag{4}$$

$$MV_v(i,j) = MV(i,j) - MV(i,j+1) \tag{5}$$

$$MV_d(i,j) = MV(i,j) - MV(i+1,j+1) \tag{6}$$

$$MV_m(i,j) = MV(i+1,j) - MV(i,j+1) \tag{7}$$

where $1 \leq i < m$ and $1 \leq j < n$ The entries in these matrices corresponds to different transition states and by observing them, we can see that most of the entries fall within the range of -5 to 5. The entries in the state matrices that are less than -5 are set to -5 and the entries that are greater than 5 are set to 5. Thus we have $S = \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$ and $|S| = 11$. This allows the reduction of the size of transition probability matrix to 11×11 without substantial compromise in the state values.

Transition probability matrices in four directions can be obtained by applying Markov-1 process on the state matrices, as,

$$T_h(p,q) = P[MV_h(i+1,j) = q | MV_h(i,j) = p)]$$

$$= \frac{\sum_{i=1}^{m-2}\sum_{j=1}^{n-1}\delta(MV_h(i,j) = p \&\& MV_h(i+1,j) = q)}{\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\delta(MV_h(i,j) = p)} \tag{8}$$

$$T_v(p,q) = P[MV_v(i,j+1) = q | MV_v(i,j) = p)]$$

$$= \frac{\sum_{i=1}^{m-1}\sum_{j=1}^{n-2}\delta(MV_v(i,j) = p \&\& MV_v(i,j+1) = q)}{\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\delta(MV_v(i,j) = p)} \tag{9}$$

$$T_d(p,q) = P[MV_d(i+1,j+1) = q | MV_d(i,j) = p)]$$

$$= \frac{\sum_{i=1}^{m-2}\sum_{j=1}^{n-2}\delta(MV_d(i,j) = p \&\& MV_d(i+1,j+1) = q)}{\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\delta(MV_d(i,j) = p)} \tag{10}$$

$$T_m(p,q) = P[MV_m(i,j+1) = q | MV_m(i+1,j) = p)]$$

$$= \frac{\sum_{i=1}^{m-2}\sum_{j=1}^{n-2}\delta(MV_m(i+1,j)=p \&\& MV_m(i,j+1)=q)}{\sum_{i=1}^{m-2}\sum_{j=1}^{n-1}\delta(MV_m(i+1,j)=p)} \tag{11}$$

where $-5 \leq p, q \leq 5$ and $\delta(x) = \begin{cases} 1 & \text{if } x \text{ is True} \\ 0 & \text{if } x \text{ is False} \end{cases}$

From these four matrices, we can calculate a single transition probability matrix by averaging them as,

$$T = \frac{T_h + T_v + T_d + Tm}{4} \tag{12}$$

From T, we can track the small and large transitions by noting the probability values. Diagonal elements in T show the probabilities for small and smooth transition while minor- diagonal elements show probabilities for large and abrupt transitions. For an authentic video, the transition from one frame to another will be smooth and the transition probability matrix will show large probability values along the diagonal and minor-diagonal shows small probability values. Hence the probability matrices will show uniform characteristics. But for fake videos, in some of the frames there will be sharp and sudden movement from the previous frame for some blocks and it will change the transition probability matrix characteristics of that pair of frames. Thus the uniformity of transition probability matrices for subsequent frames will be lost. Thus, transition probability matrix here acts as a feature vector that can classify the video as a real one or fake.

3 Implementation and Results

The proposed algorithm was tested in various authentic and fake ballistic motion videos downloaded from video sharing websites such as YouTube. First the frames at time interval t are extracted from the video to be tested. Then for each pair of adjacent frames ft at time t and fr at time $t - \Delta t$, the motion in horizontal(MV_x) and vertical(MV_y) directions is calculated using block matching algorithm without interpolation as explained in section 2.1. These two motion vectors are combined by (3) to form MV. Fig.1 shows the motion vector obtained for two sets of consecutive frames. The block size used is 8 and the search parameter used is 10 while minimum value of block size and search parameter is set as 6. As the search parameter is 10, the range of motion magnitude is from -10 to 10. Fig.1(a) is the frames of size 480×856 from a static camera video and hence the motion vector is of size 60×107. Here motion vector shows that the motion is only for the foreground objects and the background is stationary. Fig.1(b) is the frames of size 360×640 from a moving camera video and hence the motion vector is of size 45×80. Here motion vector shows the combined motion of both foreground and background.

From MV, four state matrices are calculated using (4), (5), (6) and (7) whose values ranges from -10 to 10. As mentioned in section 2.2, the state matrix values are truncated to the range of -5 to 5. Then by applying (8), (9), (10) and (11) on the state matrices, corresponding transition probability matrices

are calculated. Then they are integrated using (12) to form T. Similarly for all pairs of frames,T is calculated and their behaviour is compared by plotting the norm values against the frame number. Fig.2-Fig.4 shows frames of some of the videos tested and corresponding transition probability matrix plots. Fig.2 shows an authentic and a fake video from static camera. Fig.3 and Fig.4 shows the results of authentic videos from moving camera and fake videos from moving camera respectively.

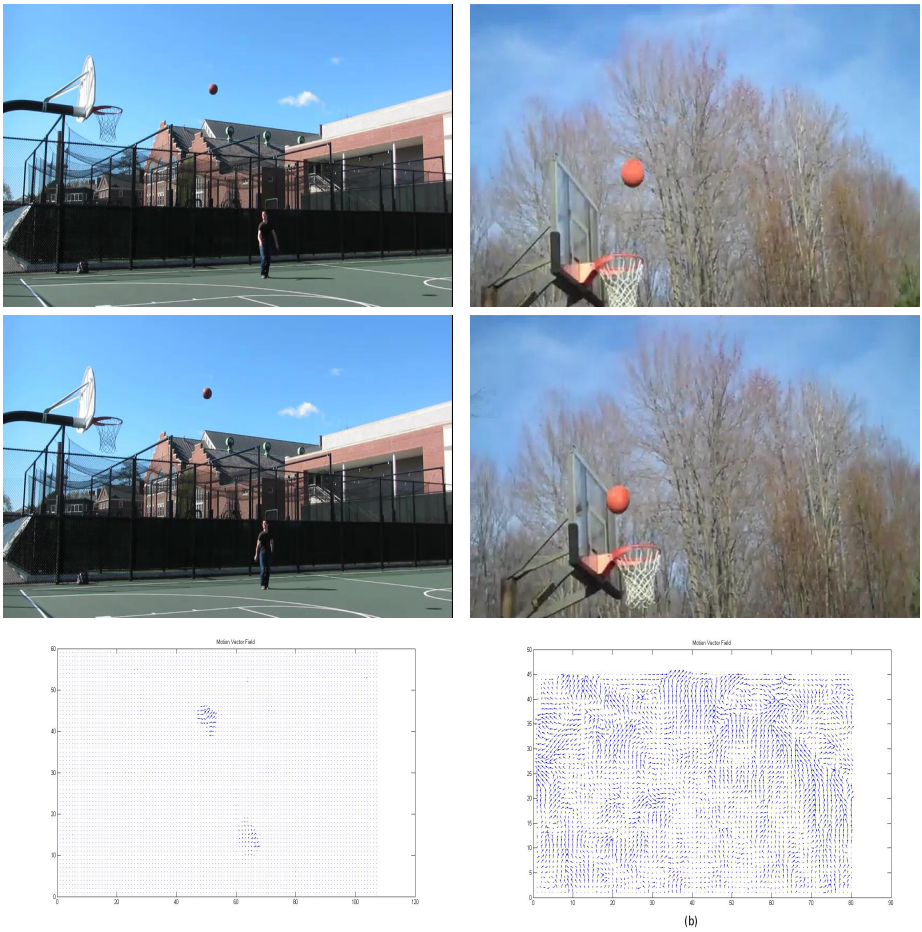


Fig. 1. Motion Vector.(a) First two rows show two successive frames from a static camera video and the third row shows corresponding motion vector field. (b). First two rows show two successive frames from a moving camera video and the third row shows corresponding motion vector field.

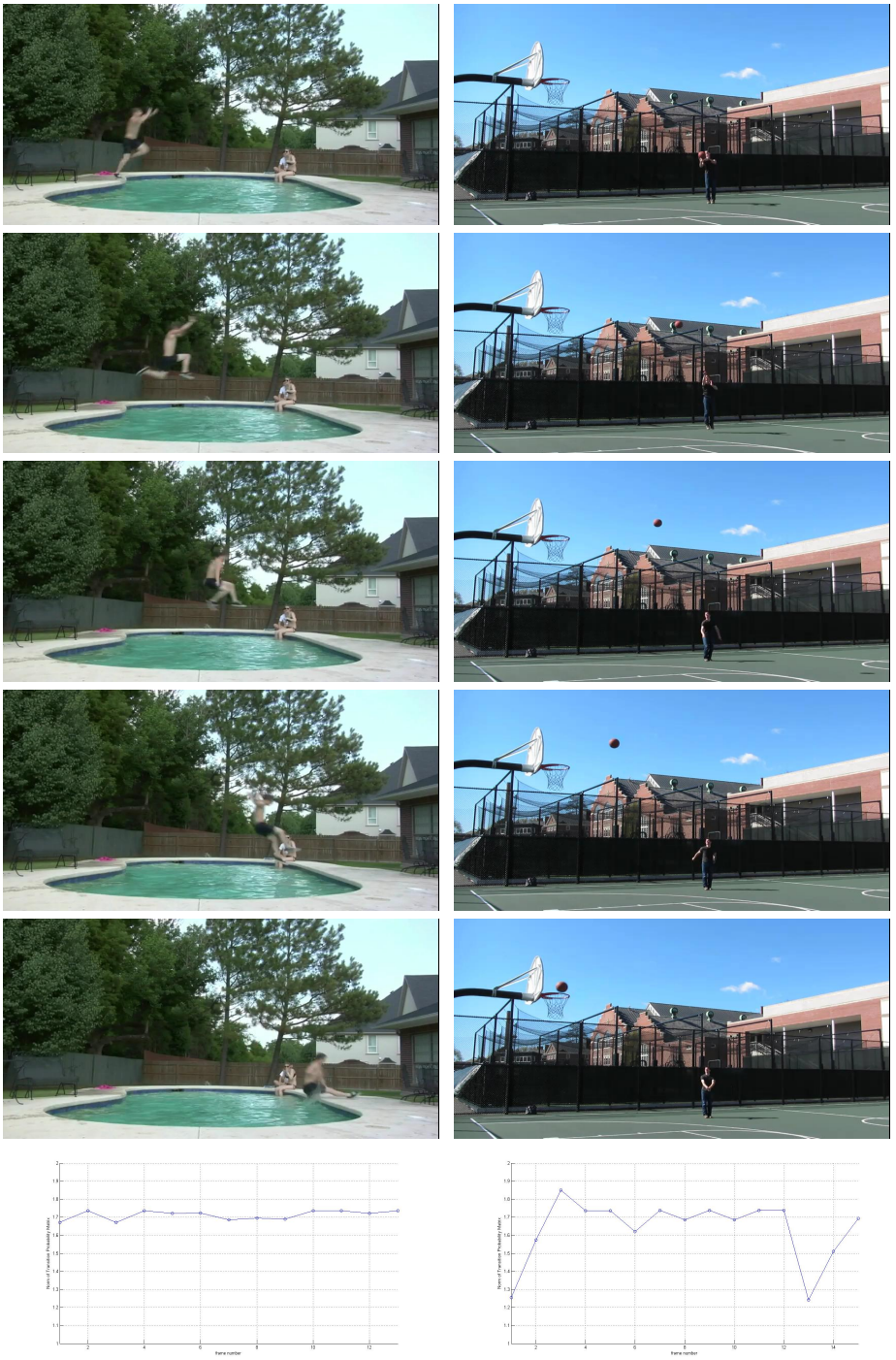


Fig. 2. First column shows an authentic video and second column shows a fake video from static camera and their corresponding transition probability matrix plots

Fig. 3. Authentic videos from moving camera and their corresponding transition probability matrix plots

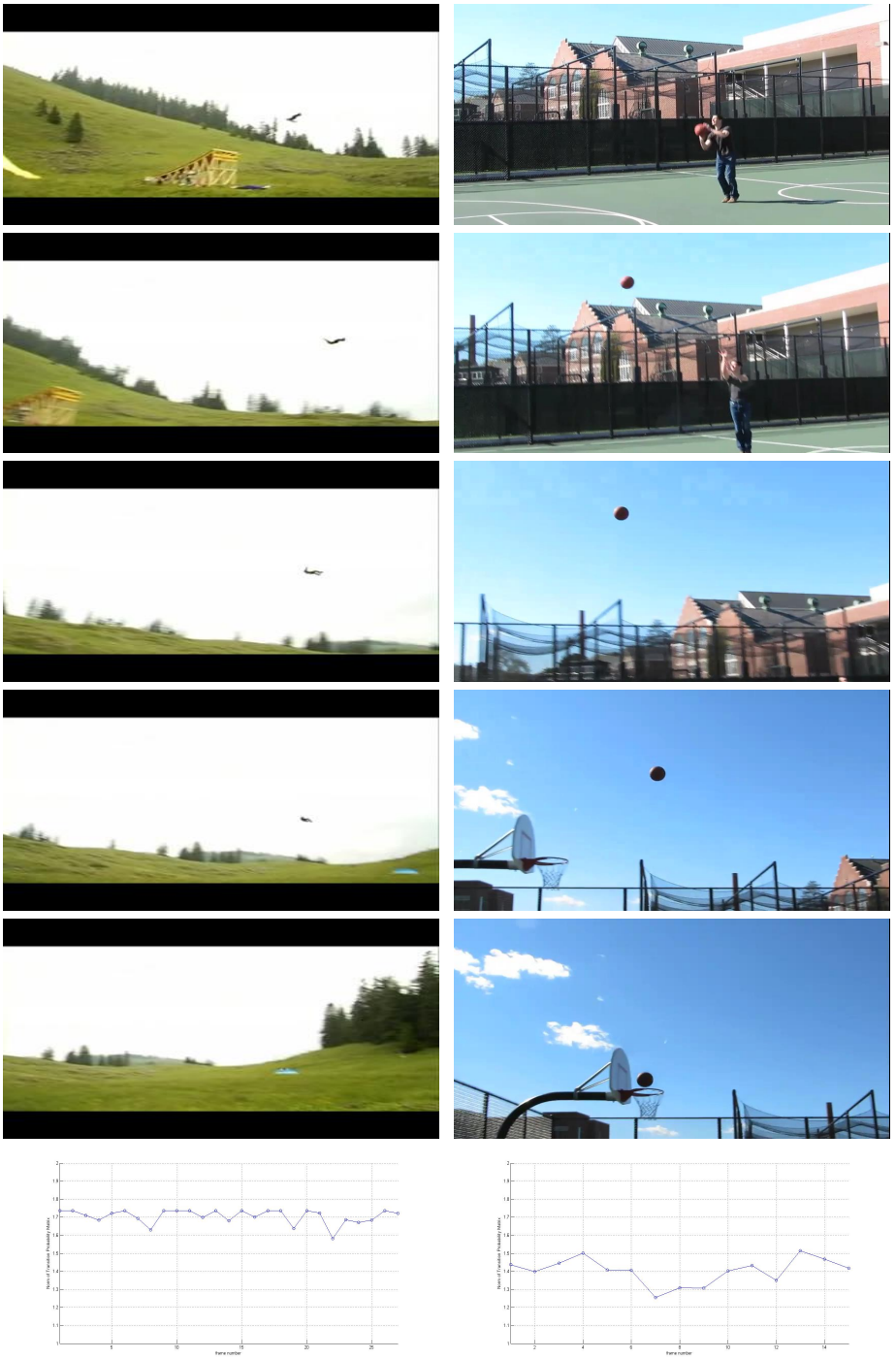


Fig. 4. Fake videos from moving camera and their corresponding transition probability matrix plots

For authentic videos, the probability value for each transition in the transition probability matrices of all pair of adjacent frames is almost the same. Hence they show a consistent behaviour which is reflected in the constant norm value. This observation holds for both static camera videos as shown in Fig.2 and for moving camera videos as shown in Fig.3. In the case of moving camera videos, camera motion plays a role in estimated motion. For obtaining good results in moving camera videos, the camera motion should be a controlled constant one. If the background is moving in an irregular random way, the motion vectors show large transitions which alter the constant behaviour of transition probability matrices and produces inaccurate results. For fake videos, the transition probability matrices of adjacent frames do not show any consistent behaviour and the plot of norm value shows zig-zag behaviour as shown in Fig.2 and Fig.4.

4 Conclusion

In this work, we have proposed a new method for detecting forgery in ballistic motion videos by motion estimation and modelling. The video is converted to frames and the motion between consecutive video frames is calculated as a motion vector using block matching algorithm. The motion vectors of all pair of frames are modelled using a Markov-1 process model and the characteristics of transition probability matrices are analysed to determine whether the video is authentic or fake. If the transition probability values remains almost constant for all pair of frames, we conclude that it is an authentic video, whereas for fake videos, the transition probabilities changes for some frames. Our method proves to be good for all videos from static camera and for videos from moving camera where the camera motion is in a controlled manner.

References

1. Wang, W., Farid, H.: Exposing digital forgeries in interlaced and de-interlaced video. IEEE Transactions on Information Forensics Security 3(2), 438–449 (2007)
2. Hsu, C.C., Hung, T.Y., Lin, C.W., Hsu, C.T.: Video forgery detection using correlation of noise residue. In: Proc. IEEE Workshop on Multimedia Signal Processing, pp. 170–174 (2008)
3. Houten, W.V., Geradts, Z.J.: Source video camera identification for multiply compressed videos originating from YouTube. Digital Investigation 6(1-2), 48–60 (2009)
4. Mondaini, N., Caldelli, R., Piva, A., Barni, M., Cappellini, V.: Detection of malevolent changes in digital video for forensic applications. In: Proc. SPIE Conf. Security, Steganography, and Watermarking of Multimedia Contents (2007)
5. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting double MPEG compression. In: Proc. ACM Multimedia and Security Workshop, pp. 37–47 (2006)
6. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting double quantization. In: Proc. ACM Multimedia and Security Workshop, pp. 39–48 (2009)
7. Chen, W., Shi, Y.Q.: Detection of double MPEG video compression using first digits statistics. In: Proc. Int. Workshop on Digital Watermarking, pp. 16–30 (2008)
8. Conotter, V., OBrien, J.F., Farid, H.: Exposing Digital Forgeries in Ballistic Motion. IEEE Trans. Inf. Forensics Security 7(1), 283–295 (2012)

9. Thormahlen, T., Broszio, H.: Voodoo Camera Tracker.: A Tool for the Integration of Virtual and Real Scenes,
http://www.digilab.uni/~hannover.de/docs/manual.html
10. Chan, S.H., Vo, D.T., Nguyen, T.Q.: Subpixel Motion Estimation Without Interpolation. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 722–725 (2010)
11. Shi, Y.Q., Chen, C., Chen, W.: A Markov Process Based Approach to Effective Attacking JPEG Steganography. In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 249–264. Springer, Heidelberg (2007)
12. Ghanbari, M.: The Cross-Search Algorithm for Motion Estimation. IEEE Transactions on Communications 38(1), 950–953 (1990)

An Extended Region Incrementing Visual Cryptography Scheme Using Unexpanded Meaningful Shares

T. Anila and M. Wilscy

Dept. of Computer Science
University of Kerala, Kariavattom,
Thiruvananthapuram, Kerala, India

Abstract. Region Incrementing Visual Cryptography (RIVC) is an important and active research area. In (2, *n*) region incrementing visual cryptographic scheme, a single secret image is divided into multiple secret regions. We need at least 2 secret shares to be superimposed to reveal the 1st secret region. By stacking more and more shares we get the entire image revealed. All the existing schemes for RIVCS suffer from the problems like color reversal, pixel expansion and low contrast. Also conventional RIVCS schemes generate noise like shares. The management of these shares is also problem, as there is no way to identify these shares. Thus we are proposing a new Extended RIVC scheme which uses the characteristics of both RIVCS and Extended VCS which adds a meaningful cover image to the shares. Our proposed method solves all the above mentioned drawbacks of the existing schemes. The main contributions in this paper are: a) No pixel expansion b)Meaningful share images c)Increased contrast d) No color reversal e) Improved Security.

Keywords: region incrementing visual cryptography, secret shares, pixel expansion, secrecy levels, color reversal.

1 Introduction

Visual Cryptography (VC) is a secret sharing scheme proposed by Naor and Shamir [1] which encodes secret messages in the form of images. The resulting encoded shares are superimposed at the receiving end to reveal the secret. The first VC scheme designed is referred as (*k,n*) VCS which encrypt a secret image into *n* transparencies(shares) such that superimposing less than *k* shares will not reveal the secret. Any *k* or more shares can be superimposed to reveal the secret.

Some of the major contributions in the area of Visual Cryptography are Probabilistic VCS [2]-[4],Visual Cryptography for General Access Structures [5] , Progressive visual cryptography[6]-[9], Extended visual cryptography for natural images[10], Halftone Visual Cryptography [11]-[12], Visual cryptography for color images[13]-[14], Tagged Visual Cryptography[15], Multi Secret VCS [16]-[17], Region incrementing visual cryptography[18]-[20]. In this paper, we are dealing with region incrementing visual cryptography (RIVCS) [18] in which a single image is considered as multiple secret regions and each secret is revealed by stacking the

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 340–349, 2013.
© Springer International Publishing Switzerland 2013

shares according to the secrecy levels assigned to them. All of the existing methods for constructing RIVCS suffer from large pixel expansion and a loss in contrast. The rest of the paper is organized as follows. Section 2 is a brief review of related works on RIVCS and the basic model of *(k,n)* RIVCS. The idea and construction methods for the proposed RIVCS are given in Section 3. Section 4 shows the experimental results and a brief conclusion is made in Section 5.

2 Related Works

Region Incrementing Visual Cryptography was introduced by R Z Wang [18] in 2009. In a *n*-level RIVCS, a secret image can be considered as an image consisting of (*n*-1) regions from which we can create *n* shares, such that superimposing more and more shares reveal the secrets according to the secrecy level.

In the original RIVCS, there exists a problem of pixel expansion and color reversal. For a binary image, color reversal is not a problem, as it contains only black and white pixel combinations. But if the color of the image itself is a secret or if the image is a greyscale image, this is a problem. Also a large pixel expansion leads to inefficiencies in the transmission.

Many works have been reported in the literature to reduce the pixel expansion. But still there is not a perfect solution for dealing with pixel expansion in RIVC. In 2012, Yang et al. proposed a *k* out of *n* region incrementing scheme [19] in which the color reversal and pixel expansion are considerably reduced. But the elimination of color reversal results in large pixel expansion and vice versa. Also, the size of basis matrices is increasing with the number of secret shares, *n*. In the same year, Shyu and Jiang also proposed an Efficient Construction for Region Incrementing Visual Cryptography [20] using linear programming. Here also the color reversal and pixel expansion exists.

2.1 Basic Model of RIVCS

In a binary VCS, the secret image consists of black and white pixels and each pixel is encoded using *m* black and white sub pixels yielding *n* shares. Each share looks like noise like images so that no secrets are revealed by superimposing less than *k* shares. If *k* or more than *k* shares are superimposed, the secrets get revealed according to the secrecy level. For each pixel, there is an encoding matrix for each security level. These matrices are called basis matrices of white (or black) pixels. They are represented as LK_r^0 (or LK_r^1) where r =1, 2. The collection of basis matrices are denoted as C_0 and C_1 and each collection consist of every possible combination of permuting the columns of LK_r^0 and LK_r^1 of all secrecy levels.

In RIVCS, the secret image is subdivided into multiple regions in such a way that any k shadow images, where $2 \leq k \leq n$, can be used to reveal the $(k-1)^{th}$ region[18]. For example, in (2, 3) RIVCS, the original image is divided into 2 secret regions as shown in Fig.1 and we get 3 shares after encoding. After encryption we get 3 secret shares. While combining any 2 shares, we get the 1st secret region revealed and by stacking all the 3 shares we get the 2nd secret region revealed.

RIVCS works as follows: Let LK_r^0 and LK_r^1 are the basis matrices of white and black pixels in the r^{th} security level. If the pixel 'p' to be processed is a white pixel, randomly choose an encoding matrix from the r^{th} level encoding matrix LK_r^0 to encode the pixel. If it is a black pixel, randomly choose an encoding matrix from the r^{th} level encoding matrix LK_r^1 . Repeat the process until all of the pixels are processed. The resulting n shares can be superimposed to get the secrets according to the secrecy levels. The column permutation is necessary because the secrets may get revealed in some shares.

While designing the basis matrices for a region incrementing scheme we need to consider the security and contrast conditions of VCS [1] . Let H (.) be the hamming weight function which indicates the total number of black pixels and V = OR (i_1, i_2 … i_k) where i_1, i_2 … i_k are the k rows of the basis matrix under consideration (LK_r^0 or LK_r^1). Then any (k, n) VCS can be considered as valid if the following security and contrast conditions are satisfied.

1. ***Contrast condition***: The number of black sub pixels in the 'OR' result of any k rows of LK_r^1 is more than the number of black sub pixels in the 'OR' result of any k rows of LK_r^0.

2. ***Security condition***: The number of black pixels in the 'OR' result of less than k rows in LK_r^1 is equal to the number of black pixels in the 'OR' result of less than k rows i_n LK_r^0.

Two important parameters of any VCS are pixel expansion and the contrast. The pixel expansion, m, can be referred as the number of sub pixels used to encode a pixel in the original image. The resulting shares are m times larger than the original image. The relative difference, α, between the black and white pixels of the superimposed image is called the contrast and it can be calculated as $\alpha = \frac{h-l}{m}$. The main aim of proposed RIVC scheme is to reduce the pixel expansion and to increase the contrast. In all the previous schemes there is an increase in pixel expansion and reduction in contrast in the reconstructed image. Also most of the previous schemes suffer from incorrect color problem where black and white pixels of the original image are reversed in the superimposed image.

For a (2,3) scheme, there are 4 basis matrices $LK_1^0, LK_2^0, LK_1^1, LK_2^1$ where $LK_1^0 = LK_2^0$ to encode the secret[18]. Any single row in LK_n^m, m=0, 1; n=1, 2 contains 2 black and 2 white pixels, so that each share appear with a uniform contrast. Stacking any 2 shares with a contrast ¼ and stacking all the 3 shares reveals the secrets with contrasts ¼ and ¼ respectively. The basis matrices are given below.

$$LK_1^0 = LK_2^0 = \begin{bmatrix} 0\,0\,1\,1 \\ 0\,1\,0\,1 \\ 0\,1\,1\,0 \end{bmatrix} \quad LK_1^1 = \begin{bmatrix} 0\,0\,1\,1 \\ 0\,0\,1\,1 \\ 0\,0\,1\,1 \end{bmatrix} \quad LK_2^1 = \begin{bmatrix} 0\,0\,1\,1 \\ 0\,1\,0\,1 \\ 1\,0\,0\,1 \end{bmatrix}$$

Here pixel expansion m=4 and Contrast $\alpha = \frac{1}{4}, (\frac{1}{4}, \frac{1}{4})$ for 1^{st} , (1^{st}, 2^{nd}) security levels.

But Wang's scheme suffers from pixel expansion and incorrect color problem. Later, Yang et. al[19] proposed a general construction method for (k,n) RIVCS, where it requires at least k shares to be superimposed to get one secret. They proposed

two construction methods for (2, 3) RIVCS. One method solves incorrect color problem and other solves pixel expansion of 1st method, but with a color reversal. Method 1 has a pixel expansion of 6 and Method II has a pixel expansion of 4. So both of these methods are not a good solution for incorrect color problem and pixel expansion.

In this paper, we propose an efficient (2,3) RIVC scheme which solves all these problems. Also it uses the extended VCS[21] method to add a cover image to make the shares meaningful. This solves the problem of share management. The following section discusses the details about our proposed method.

3 Proposed Extended RIVCS

In the conventional RIVC schemes, pixel by pixel processing is done to encode the secrets in the shares. In the proposed (2, *n*) RIVCS method, in order to make the pixel expansion constant, a group of pixels will be processed at a time. 4 neighbouring pixels are replaced by 4 pixels in each share. If the hamming weight of all the 4 pixels is greater than a particular threshold *c*, it is considered as a black pixel and use the black pixel encoding basis matrix to encode the group of pixels. Otherwise consider it as a white pixel and use the corresponding white pixel basis matrix to encode the group of pixels.

3.1 Proposed (2,3) RIVCS

In (2,3) RIVCS ,there will be 2 secret levels and 3 shares for a secret image. We need at least 2 shares to reveal the 1st secret region and all the 3 shares to be superimposed together to get the 2nd secret. Let the basis matrices for encoding the white pixel of the r^{th} security level (r=1, 2) be LK_r^0 and that of black pixel is LK_r^1 both having an order $n \times m$, where *n*=3 and *m*=4. To maintain the security and contrast condition use the permuted columns of LK_r^0 and LK_r^1 for all *r*, where $r \in [1, n-1]$. Also the basis matrices for the black pixel group should be the same. That is; $LK_1^1 = LK_2^1$. This is to correct the non uniformity between the secret levels after superimposing the shares. In order to improve the security of the secret shares and we can add different gray scale images to the noisy secret shares. Also this allows the dealers to identify the order of shares. This yields 3 meaningful grayscale shares so that no receiver gets a clue that another secret is hidden inside the shares. By superimposing any two shares, the 1st security level region gets revealed and by superimposing all the shares, all the secret regions get revealed.

Algorithm:

1. Divide the original image into two secret regions by assigning the secrecy levels r = 1,2 to each group of 4 neighbouring pixels.
2. For each security region *r* do the following steps.
    a. For each pixel group (*p*) calculate the hamming weight of black pixels and let it be *H*(*p*).

i. If $H(p) >= c$, consider p as set of black pixels. Also encode p using LK_r^1.
ii. Otherwise, treat the group p as a set of white pixels and encode p using LK_r^0.

3. Select three gray scale images, with the same size of the original secret image, as cover images for the 3 noise like shares obtained from step 2.
4. For each share do the following:
   a. For each pixel of the secret share, take the corresponding pixel from one of the gray scale image and perform AND operation.
   b. The resulting gray scale image retain the gray values if the corresponding pixel value of the secret share is 1, otherwise the pixel value is set as zero.
5. The resulting shares are 3 noisy gray scale images which are entirely different from encoded secret image.

The basis matrices for black and white pixel group are given by:

$$LK_1^0 = \begin{bmatrix} 0\,0\,1\,1 \\ 0\,0\,1\,1 \\ 1\,1\,0\,0 \end{bmatrix} \quad LK_2^0 = \begin{bmatrix} 1\,0\,0\,1 \\ 1\,0\,1\,0 \\ 0\,0\,1\,1 \end{bmatrix} LK_1^1 = LK_2^1 = \begin{bmatrix} 1\,0\,0\,1 \\ 1\,0\,1\,0 \\ 1\,1\,0\,0 \end{bmatrix}$$

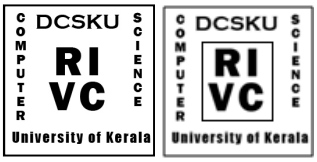
Let $\left|\mathrm{LK}_\mathrm{r}^0\right|$ be the number of columns of r-th security level where r=1,2. Since $\left|\mathrm{LK}_\mathrm{r}^0\right| = \left|\mathrm{LK}_\mathrm{r}^1\right|$, no secret is revealed and every row has equal black and white pixels resulting in noise like shares. In this paper, the threshold applied is 3. By combining any two shadows we will get 2 black and 2 white pixels or 4 black and 0 white pixels in LK_1^0 and 3 black and 1 white pixels in LK_1^1 yielding the 1st security level region with a contrast of 1. No 2nd level secret is revealed as it has 3 white and 1 black pixels in LK_2^0 and LK_2^1 while combining any 2 shadows. Thus it holds the security condition . Similarly, by combining all the 3 shares, we get 3 black and 1 white pixels in LK_2^0 and 4 black and 0 white pixels in LK_2^1 .This reveals the 1st and 2nd level secret with contrast 1.

From the above discussion we can conclude that it is possible to implement (2, 3) RIVCS, without pixel expansion using meaningful shares. Also we need not require complex basis matrices in order to encode the secret messages. Compared to the previous RIVC schemes, our scheme has no pixel expansion, correct color and increased contrast without compromising the security. In all other schemes, the pixel expansion, low contrast and color reversal result in a low contrast and poor visual quality.

4 Experimental Results

In order to prove the feasibility of our method some of the experimental results are provided. All of the schemes are implemented using Matlab. Fig 1(a) shows the original secret image to be encoded. In Fig 1(b) the letters outside the rectangular box belong to region 1 having secrecy level 1 and the letters inside the rectangular region

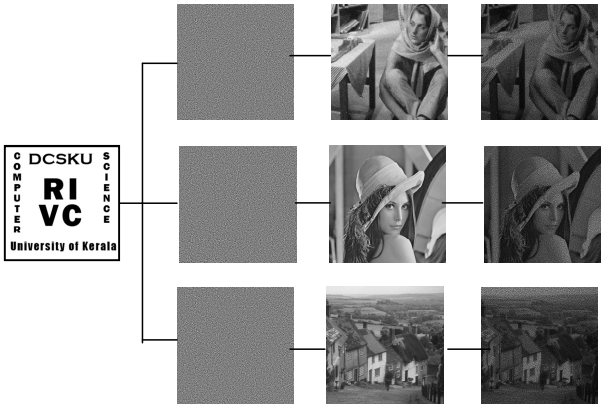
belong to region 2 having secrecy level 2. In (2,3) RIVCS, we get 3 meaningful shares after encoding. After superimposing any 2 shares the letters having the security level 1 is revealed. By stacking all the 3 shares, the entire image gets visible. The stacked results shows an entirely different image from the secret shares. This shows the high security for the original secret image.



(a) (b)

Fig. 1. Secrecy level decomposition of (2,3) RIVCS (a)Original secret image (b) 1st and 2nd level of secret

Fig.2(a) shows the intermediate noise like shares before embedding the cover images into the shares. Fig.2(b) is the three different images selected as cover images. Fig.2(c) represents the resulting share images after embedding the cover images into the secret shares. All of the images are the same size as the original image. This means that there is no pixel expansion happens to any of the secret shares. The resulting shares are the noise added cover images which does not have any physical similarity with the encoded secret image. Fig.3 is the results of 1st level of superimposing the results. When any of the two meaningful shares are superimposed, the 1st secret region with secrecy level 1 is revealed. No trace of the cover images is seen in the resulting superimposed image.Fig.4 is the actual output image when superimposing all the shares. The resulting image shows correct color, no pixel expansion and increased contrast compared to other schemes.



(a) (b) (c)

Fig. 2. Proposed (2,3) RIVCS (a) initial noise like shares (b) cover images (c) final meaningful secret shares

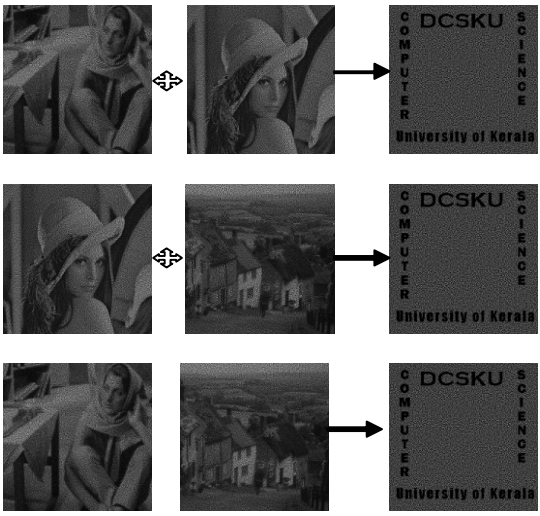


Fig. 3. Proposed (2,3) RIVCS – Stacking any two shares reveal 1st secret level region

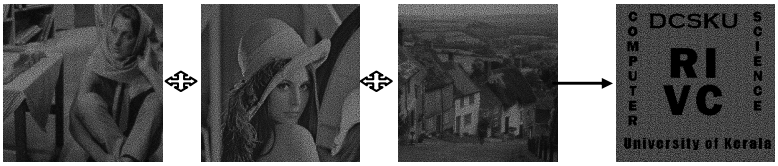


Fig. 4. Proposed (2,3) RIVCS – Stacking all the three shares reveal entire secret image

Fig.5 is the comparison of various (2, 3) RIVC schemes. All the existing RIVC schemes suffer from large pixel expansion, low contrast and color reversal. Fig 5(a) is the original secret image to be encoded. Fig 5(b) is the result of Wang's (2, 3) scheme which reverses the color of the 1st secret region and also the size of the superimposed image is 4 times bigger than the original image. Fig 5(c) and (d) are the result of Yang et. al's *k* out of *n* RIVCS. They have proposed two methods for (2, 3) RIVCS . 1st method reveals the superimposed image with correct colors but with an increase in pixel expansion (m=6). Method II reduces the pixel expansion to 4 but with a color reversal. Fig 5(e) is the superimposed image of our proposed RIVCS with no pixel expansion. Our proposed method shows that the superimposed image is exactly the same size as the original image without any difference in the color of any of the secret regions. This means that our proposed method is more efficient compared to other existing schemes.

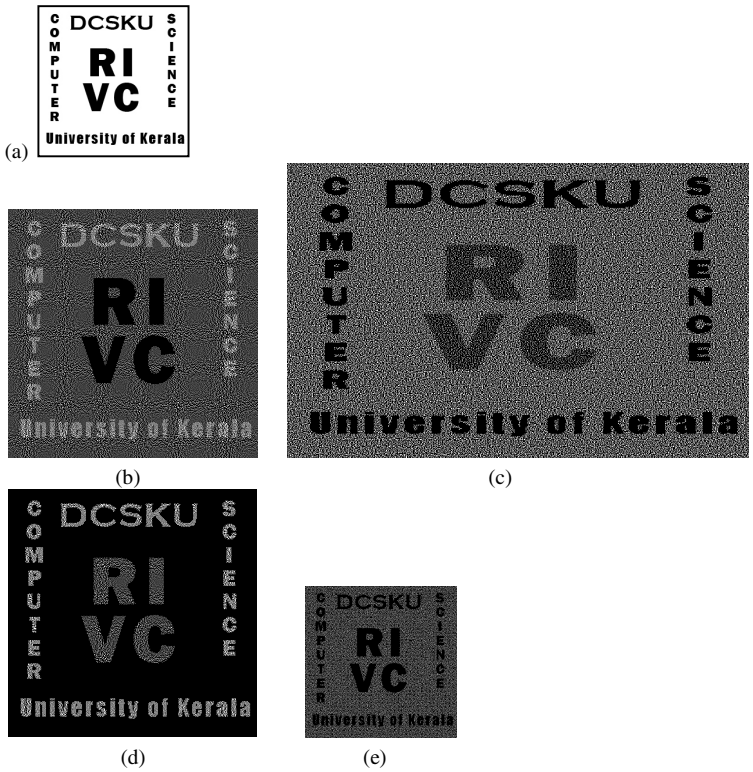


Fig. 5. Results of various (2,3) RIVCS (a)Original secret image (b) Result of Wang's scheme (m=4) (c)Result of Yang et,al's *k* out of *n* RIVCS(method 1, m=6) (d) Result of Yang et,al's *k* out of *n* RIVCS(method 2, m=4) (e)Result of proposed method

Table 1 is the comparison of pixel expansion of different schemes. Wang's RIVCS and Yang et al's Scheme II has pixel expansion 4 and Yang et al's Scheme I shows a pixel expansion 6. But our proposed scheme has no pixel expansion (m=1). That is , the original image and the resulting superimposed image are of the same size. Another performance measure of any VCS is the contrast. The experimental results show that our proposed method has better contrast compared to all other existing RIVCS. Table 2 is the comparison of contrasts of our proposed method with other schemes. Our scheme shows an increase in the contrast for both of the secrecy levels and the contrast is calculated as 1. Other schemes have lesser contrast compared to our proposed scheme. Also our scheme shows a uniform contrast for both secrecy levels. The non uniformity and color reversal of Wang's scheme is completely eliminated in the proposed scheme.

Table 1. Comparison of Pixel Expansion of (2,3) RIVCS

Method	Pixel Expansion
Wang's scheme	4
Yang et al's scheme I	6
Yang et al's scheme II	4
Our proposed scheme	1

Table 2. Comparison of Contrast While Stacking the Shares

RIVC	Security level	Contrast					
		Wang's Scheme		Yang et al's Scheme		Proposed Scheme	
		No: of superimposed shares					
		2	3	2	3	2	3
(2,3)	1st	¼	¼	¼	½	1	1
	2nd	-	¼	-	¼	-	1

5 Conclusion

In this paper, an Extended Region Incrementing Visual Cryptographic method is proposed for hiding secrets in meaningful share images without pixel expansion. In the traditional visual cryptographic system noise like share images are being used. We can give additional security by adding cover images to these shares so that nobody gets a clue that there is a second secret inside the secret share. Moreover the share management problem in the conventional VCS is solved here. Also the incorrect color problem of all previous schemes is completely eliminated. Our results show a vast difference in the pixel expansion without compromising the security compared to other existing schemes.

References

1. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
2. Yang, C.N.: New visual secret sharing schemes using probabilistic method. Pattern Recognit. Lett. 25(4), 481–494 (2004)
3. Cimato, S., De Prisco, R., De Santis, A.: Probabilistic visual cryptography schemes. Comput. J. 49(1), 97–107 (2006)
4. Wang, D., Yi, F., Li, X.: Probabilistic visual secret sharing schemes for grey-scale images and color images. Inform. Sci. 181(11), 2189–2208 (2011)
5. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Visual Cryptography for General Access Structures. Information and Computation 129(2), 86–106 (1996)
6. Fang, W.P., Lin, J.C.: Progressive viewing and sharing of sensitive images. Patt. Recog. Image Anal. 16(4), 638–642 (2006)

7. Fang, W.P.: Multilayer progressive secret image sharing. In: Proc. 7th WSEAS, pp. 112–116 (2007)
8. Fang, W.P.: Friendly progressive visual secret sharing. Patt. Recog. 41(4), 1410–1414 (2008)
9. Hou, Y.-C., Quan, Z.-Y.: Progressive Visual Cryptography with Unexpanded Shares. IEEE Transactions On Circuits And Systems For Video Technology 21(11) (November2011)
10. Nakajima, M., Yamaguchi, Y.: Enhancing registration tolerance of extended visual cryptography for natural images. J. Electron. Imag. 13(3), 654–662 (2004)
11. Zhou, Z., Arce, G.R., Crescenzo, G.D.: Halftone visual cryptography. IEEE Trans. Image Process. 15(8), 2441–2453 (2006)
12. Wang, Z., Arce, G.R., Crescenzo, G.D.: Halftone visual cryptography via error diffusion. IEEE Trans. Image Process. 4(3), 383–396 (2009)
13. Hou, Y.C.: Visual cryptography for color images. Patt.Recognit. 36(7), 1619–1629 (2003)
14. Jin, D., Yan, W.Q., Kankanhalli, M.S.: Progressive color visual cryptography. J. Electron. Imag. 15(3), 033019:1–033019:13 (2005)
15. Wang, R.-Z., Hsu, S.-F.: Tagged Visual Cryptography. IEEE Signal Process. Lett. 18(11) (November 2011)
16. Shyu, S.J., Huang, S.Y., Lee, Y.K., Wang, R.Z., Chen, K.: Sharing multiple secrets in visual cryptography. Pattern Recognit. 40, 3633–3651 (2007)
17. Feng, J.B., Wu, H.C., Tsai, C.S., Chang, Y.F., Chu, Y.P.: Visual secret sharing for multiple secrets. Pattern Recognit. 41(12), 3572–3581 (2008)
18. Wang, R.Z.: Region incrementing visual cryptography. IEEE Signal Process. Lett. 16(8), 659–662 (2009)
19. Yang, C.-N., Shih, H.-W., Wu, C.-C., Harn, L.: K Out of n Region Incrementing Scheme in Visual Cryptography. IEEE Trans. on Circuits and Systems for Video Technology 22(5) (May 2012)
20. Shyu, S.J., Jiang, H.-W.: Efficient Construction for Region Incrementing Visual Cryptography. IEEE Trans.on Circuits and Systems for Video Technology 22(5) (May 2012)
21. Ateniese, G., Blundo, C., Santis, A.D., Stinson, D.R.: Extended capabilities for visual cryptography. Theor. Comput. Sci. 250, 143–161 (2001)

Online Signature Verification Based on Recursive Subset Training

D.S. Guru[1], K.S. Manjunatha[1], and S. Manjunath[2]

[1] Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore – 570 006, Karnataka, India
dsg@compsci.uni-mysore.ac.in, kowshik.manjunath@gmail.com
[2] Department of Studies in Computer Science, JSS College of Arts, Commerce and Science, Mysore – 570 025, Karnataka, India
manju_uom@yahoo.co.in

Abstract. In this paper, a novel approach has been proposed for online signature verification based on recursive subset training. Our approach is based on estimating the Equal Error Rate (EER) of the entire system and then splitting the entire data set into two subsets based on the EER of the system. The two subsets includes writers whose individual EER is more than the EER of the system and writers whose EER is less than the EER of the system. This procedure is recursively repeated until writer level parameters are decided. Unlike other verification models where same features are used for all writers, our approach is based on identifying writer dependent features and also writer dependent thresholds. Initially, writer dependent features are selected using a suitable feature selection method. Signatures are clustered using Fuzzy C means and represented in the form of interval valued symbolic feature vector. Signature verification is done based on the selected representation and the EER of system is calculated. Once the EER of the system is estimated, our method is based on estimating the EER of individual writers and splitting the dataset into subsets and estimating the EER of each of the subset separately. This process of splitting the dataset into subset and treating each of the subsystem separately is repeated until the individual writer thresholds and features are identified. We conducted experiments on MCYT-DB1 to show the effectiveness of our novel approach.

Keywords: Subset recursive training, writer dependent parameters, Feature selection, Fuzzy C means, Symbolic feature vector.

1 Introduction

Due to the enormous growth of internet, biometric based authentication has attained greater significance as a secured means of authentication during the last few decades. Biometric authentication is based on either physiological traits such as face, iris, finger prints or behavioral traits such as signature, gait, voice of an individual. Among all the behavioral traits, signature has widest social acceptance for identity authentication [1].

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 350–361, 2013.
© Springer International Publishing Switzerland 2013

Depending on an acquisition method, signature verification can be either offline or online [2]. In an offline mode, also known as static mode, a signature is acquired from documents and digitized. Here no special hardware is required for signature acquisition and verification is done based on signature shape only. Application of offline signature verification can be found in banks for cheque validation, validation of many legal and financial documents etc. On the other hand, in an online mode, also known as dynamic mode, a signature is acquired by means of a special hardware such as digital tablet or smart pens which can capture additional dynamic features such as pressure, speed of signing, along with signature shape. Due to the availability of additional features, it is difficult for a forger to replicate both signature shape and dynamic features and hence it is more reliable than the offline mode.

Online signature verification methods are of two types depending on the type of features used for verification namely function based features and parameter based features [2][3]. In a function based approach, signature is represented by means of a time function of dynamic signing process such as pressure, velocity, acceleration, position trajectory [4][5]. During verification the time functions of a test signature and a reference signature are matched. In a parameter based approach, a signature is represented by means of a feature vector consisting of feature values extracted from specific points in the signature and during verification, parameters of a test and a reference signatures are compared to establish the authenticity of the test signature. Even though the enrollment size and matching time is less in case of parameter based approach compared to function based, error rate is generally high in case of parametric based approach [6]. Number of features have been proposed in the literature for online signature verification which have been categorized into local and global [3][2]. Local features, are extracted from specific point in the signature and global features correspond to whole signature or major part of the signature. Some of the local features for on-line signatures are curvature change, azimuth, pen elevation, pressure, speed etc., while the global features include signature writing time, number of strokes, average speed etc. [2][5][3]. Signature verification based on global features can be found in the work of [2][7]. Similarly signature verification based on local features can be found in the work of [3]. Attempts have been made on the usage of both local and global feature for signature verification [5].

During verification, authenticity of a query signature is determined based on the similarity between query signature and corresponding reference signatures available in the knowledgebase. Similarity between a test and a reference signatures are estimated based on suitable matching techniques. Different matching techniques proposed for signature verification can be categorized into three groups i) Template matching where query and reference signatures are compared for best matching components. DTW is the most popular template matching technique for signature matching [2][8][9]. ii) Statistical approach where the comparison is based on probability estimation. Different matching techniques considered for signature verification based on statistical approach are HMM [4][10], SVM [11][12], Neural Network [13][14] iii) Structural approach which is based on syntactic representation where signature are described through primitives and are compared through graph or string matching [15][16].

Based on the literature survey, it is clear that almost all the verification models proposed in the literature are writer independent in the sense that, different parameters like number of features to be fixed for each writer for enrollments and similarity threshold are fixed globally without any consideration at writer level. Here every individual is characterized by the same features. Few attempts have been made towards the applicability of writer dependent threshold [2][7]. But these work rely on the usage of same features for all writers. In the work of [17], an online signature verification model based on writer dependent features is proposed. Here even though feature selected are writer dependent, feature dimension is not writer dependent. That means every writer is characterized by the same number of features. But in reality fixing different thresholds for different writers and also selecting different number of features for different writers is more effective for signature verification than using common features for all writers. In this paper we propose a novel approach for online signature verification which exploits the concept of writer dependent parameters. Our work is based on the novel concept of repeated subset training to achieve writer dependency.

Initially writer dependent features are extracted using a feature selection method. After writer dependent features are selected, signature samples are clustered based on selected features and each cluster is represented in the form of interval valued symbolic feature vector [7] for assimilating intra-class variation. Signature verification based on symbolic representation is presented and EER of the whole system is estimated. Depending on the EER of the whole system, the data set is split into two subsets, one consisting of writers whose individual EER is more than the EER of the whole system and another consisting of writers whose individual EER is less than the EER of the whole system. We treat each of these subsets separately and reapply feature selection, clustering and verification and calculate the EER of the subsets separately. This process is repeated till individual writer level is reached. During verification, corresponding thresholds and the number of features to be fixed for each writer is used to decide the authenticity of the test signature.

The major contribution of this work is the introduction of novel concept of subset training for achieving writer dependency in online signature verification. The paper is organized as follows: In section 2, we discuss in details different stages of the proposed model. In section 3, we discuss details of experimentation, dataset used and results obtained. Comparative analysis is presented in section 4 and finally conclusions are drawn in section 5.

2 Proposed Model

The proposed model consists of the following stages

1. Writer dependent feature selection
2. Symbolic representation based on clustering
3. Signature verification
4. Recursive subset training for fixing up writer dependent parameters.

2.1 Writer Dependent Feature Selection

Feature selection is a process of finding a subset of features for enhancing the performance of a classifier. The primary benefits for feature selection includes a considerable reduction of training time, memory requirements, better data understanding and visualization. Feature selection aims to identify most relevant features to improve the performance of machine learning models [18]. In contrast to feature transformation techniques like PCA and LDA where features are transformed from one domain to another, feature selection preserves the semantics of original features without any feature transformation.

Feature selection algorithms are categorized into two types ; filter and wrapper methods. In the filter method, subset of features are selected based on certain score computed for each feature independent of any learning algorithm. In the wrapper method, feature selection process is wrapped around a learning algorithm, whose performance is used as a basis for selecting a subset of features. Even though wrapper method yields an optimal subset, it is time consuming as it involves searching all possible subsets to arrive at an optimal subset [19][20][21][18]. Further, depending on presence or absence of class labels, feature selection methods can be categorized into supervised or unsupervised. Feature selection for unsupervised data is challenging due to the non availability of any class labels. In an unsupervised feature selection, selection of features is based on the computation of score for each feature and selecting a subset of features based on their score. In this work, we have adapted an unsupervised filter based feature selection method [22] suitable for multi cluster data. Here the score computed for selecting a feature indicates the ability of the feature in preserving a cluster structure. The details can be found in [22].

In our proposed model, signature data set of dimension $N*M*P$ where N is the number of writers, M is the number of training samples and P is the number of features is decomposed into N feature matrices of size $M*P$. The multi cluster feature selection method explained above is applied on each of these feature matrices representing an individual writer separately. It results in the reduction of dimension of feature matrix of a writer to a size $M*d$ where d is the number of features selected $(d<P)$. The indices of the corresponding features selected are also stored in the knowledgebase and it varies from a writer to a writer. These indices are stored in the database for later usage during verification.

2.2 Symbolic Representation Based on Clustering

In order to have multiple representatives for each writer, signatures are clustered based on the features selected. In our work we have adapted Fuzzy C means as it is independent of any data distribution.

One of the major challenges in signature verification is to preserve intra-class variation, which is very common in signatures. This is due to the fact that signature is a behavioural biometric which depends on the physical and mental state of the signer and there will be a lot of variations in the signature samples of a same writer. For

effective capturing of this intra-class variation, symbolic representation [7] has been proposed which is not only effective in preserving intra-class variation but also minimizes the number of feature vectors to be stored in the knowledgebase. Instead of storing every signature sample of every class in the knowledgebase, symbolic representation results in the creation of a single feature vector for each class. After clustering of signatures based on the selected writer dependent features, Symbolic feature vector for each cluster is created as follows [23]. Let $\{S_1, S_2, ..., S_n\}$ be n signature samples of a cluster C_j, $j = 1, 2, ..., K$ where K is the number of clusters in each class. Let $\{f_{j1}, f_{j2}, ..., f_{jd}\}$ be the feature vector representing the cluster C_j where d is the number of selected features. Let M_{jk}, $k = 1, 2, ..., d$ and σ_{jk}, $k = 1, 2, ...d$ be the mean and standard deviation of k^{th} feature of the cluster C_j i.e

$$M_{jk} = \frac{1}{n}\sum_{i=1}^{n} f_{ik} \quad \text{and} \quad \sigma_{jk} = \left[\frac{1}{n}\sum_{1}^{n}\left(f_{ik} - \mu_{jk}\right)^2\right]^{\frac{1}{2}} \tag{1}$$

After estimating the mean and standard deviation of the k[th] feature of the cluster C_j, the k[th] feature of the cluster C_j is represented in the form of interval-valued feature as $\left[f_{jk}^{-}, f_{jk}^{+}\right]$, where

$$f_{jk}^{-} = M_{jk} - \alpha\sigma_{jk} \quad \text{and} \quad f_{jk}^{+} = M_{jk} + \alpha\sigma_{jk} \text{ for some scalar } \alpha. \tag{2}$$

The interval $\left[f_{jk}^{-}, f_{jk}^{+}\right]$ represents the lower and upper limits of the k^{th} feature value of a signature cluster in the knowledgebase. In general each of the d features selected is represented in the form of an interval-valued feature. For example, the reference signature for the cluster C_j is thus formed as

$$RFC_J = \left\{\left[f_{j1}^{-}, f_{j1}^{+}\right], \left[f_{j2}^{-}, f_{j2}^{+}\right] ... \left[f_{jd}^{-}, f_{jd}^{+}\right]\right\}, \qquad j = 1, 2, .. K \tag{3}$$

where K is the number of clusters in each signature class. Instead of storing every sample of every cluster, it is sufficient store this interval valued symbolic feature vector of the reference signature as a representative of the entire cluster.

If there are K clusters formed for each individual writer and N is the number of writers then we have totally NK number of reference signatures in the knowledgebase instead of Nn ($> NK$) number of signatures.

2.3 Signature Verification

Once the symbolic feature vector is created for each cluster, test signature's authenticity is decided by comparing the feature of the test signature with

corresponding interval-valued feature of reference signature. During verification a query signature represented in the form of p dimensional feature vector $F_q = \{f_{q1}, f_{q2}, .. f_{qP}\}$ is considered where every feature of the test signature is crisp in nature while the corresponding feature of the reference signature is interval valued. Out of P features of query signature only d features are compared with corresponding d interval valued features of reference signatures. The indices of all the d features to be compared are available in the knowledgebase. If a feature of query signature lies within corresponding interval-valued feature of reference signature, then the acceptance count is incremented by 1. If the total acceptance count is greater than the predefined threshold, then the test signature is accepted as genuine else, it is considered as forgery.

The acceptance count is defined to be

$$A_c = \sum_{i=1}^{d} C\left(f_{ti}, \left[f_{ji}^{-}, f_{ji}^{+} \right] \right) \tag{4}$$

where

$$C\left(f_{ti}, \left[f_{ji}^{-}, f_{ji}^{+} \right] \right) = \begin{cases} 1 \; if \left(f_{ti} \geq f_{ji}^{-} \; and \; f_{ti} \leq f_{ji}^{+} \right) \\ 0 \; otherwise \end{cases}$$

Any biometric system can results in two types of error type-1 error also known as False Acceptance Rate (FAR) which denotes the number of forgery signatures which are wrongly labeled as genuine and type-2 error also known as False Rejection Rate (FRR) which denotes the percentage of genuine signature which are labeled as forgery signature. The point on a ROC curve where FAR and FRR are equal is called Equal error rate (EER).

2.4 Recursive Subset Training

Once the EER of the entire system is calculated based on the number of features selected, EER of all the writers are estimated by fixing the values of number of features selected and similarity threshold. After calculating the EER of individual writers, the whole dataset is split into two disjoint subsets with first subset consisting of writers whose individual EER is more than the EER of the system and the second subset consisting of writers whose individual EER is less than the EER of the system. Once the whole system is split into subsets, the retraining of the system starting from writer dependent feature selection and up-to signature verification stage explained above to recalculate the EER for the split subset of writers is recursively repeated. It shall be noted that all the writers of subset 1 are with a different set of features when compared to the writers of subset 2. It is also clear that the cut off thresholds for decision making also vary. This process of finding EER of the system and then splitting into two subsets based on the EER of individual writer is recursively repeated till each of the subset will be of size representing one particular writer. By

doing this way, we eventually end up in fixing up different thresholds and different set of features for each writer depending on the individual characteristics. Thereby, we build up a system based on writer dependent parameters which is first in its kind in literature.

3 Experimentation and Results

Details of dataset used, experimentation conducted and results obtained are discussed in this section.

3.1 Experimentation

Dataset: To validate our model we conducted experiments on MCYT-100 online signature dataset which consists signature of 100 signers with 25 original and 25 skilled forgeries from each writer. Here every writer is characterized by 100 global features. The details of these 100 global features of online signatures can be found in the work of (Fierrez et al, 2005). Signature dataset is divided into training and testing set. In our work we have considered 20 genuine signatures for training purpose. All the remaining genuine signatures and all the 25 skilled forgeries were used for testing purpose.

Experimental Setup: In our work, we first conducted feature selection experiments on DB1 for varying number of features from 5 to 75 in step of 5. We conducted 20 trials by randomly selecting different training signatures in each case and the average of 20 trials is considered as EER of the system. We also noted down the corresponding values of number of features to be selected and the similarity threshold. Once the EER of the system is calculated by considering all writers, we estimated the EER of individual writers by fixing the values of features to be selected and similarity threshold. We sorted the EER of individual writer in the decreasing order and split the number of writers into two disjoint subsets with one subset containing writers whose EER is above the global EER and other subset containing writers whose EER is below the global EER. We then considered each of the two subsets separately and conducted verification experiments for varying number of features and estimated the EER of each of these subsets. This process of estimating the EER of the whole system, splitting data set into subsets based on the EER obtained and conducting feature selection and verification on each of the subset is repeated recursively till number of features and similarity threshold to be fixed for each writer is decided. Once the writer dependent parameters are achieved, these values are used to authenticate a query signature of the claimed writer.

3.2 Experimental Results

In this section we present the experimental results of our method. In the first stage when all the 100 writers are considered we obtained an EER of 5.06 for similarity threshold = 0.50 for 50 number of features. In the second stage, we calculated the

EER of individual writers based on subset training methodology and recursively split the datasets into two disjoint subsets based on the global EER. Table-1 shows the features to be fixed for each writer, similarity threshold and corresponding EER. Figure-1 shows the hierarchical structure of split of dataset for first three levels. Root node at each level represents the number of writers and its right child represents number of writers whose EER is above the EER of number of writers in its parent node and its left child represents the number of writers whose EER is less than the EER of number of writers in its parent node. Data given at the left and right of each node represent threshold fixed and EER achieved respectively when the system is trained with the number of writer specified at a given node. The EER is reduced to 2.61 when the average of individual writer at the last level where individual writer's threshold and features are decided. The EER of the system is reduced considerably and for more than 30 writers the EER is 0.

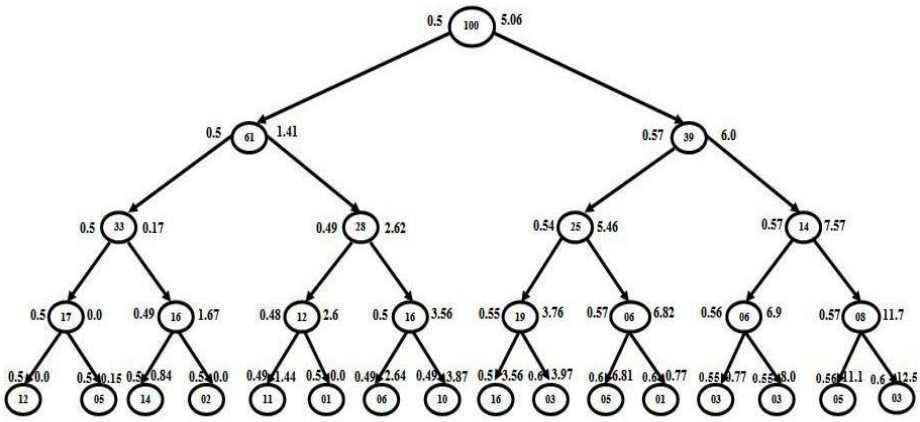


Fig. 1. Heirarchical representation of writers split for four level

4 Comparative Analysis

Comparing the performance of different verification model is difficult due to variations in dataset used, number of training and testing samples considered and also performance measures adapted. Most of the researcher have reported their results based on their own data set. The two datasets commonly used by most of the researcher for validating their model are MCYT and SVC online signature datasets. In this work, we have used MCYT-100 consisting of online signatures of 100 writers for validating our model. In this work, we compare the performance of our model with other existing models for online signature verification based on dataset used, training and testing sample size and verification results. Table-2 shows the comparative analysis of our model with other existing models. In addition we compared our

Table 1. Features Number, Similarity Threshold and EER of Individual Writers

Writer #	Features	Thresh	EER	Writer #	Features	Thresh	EER
1	60	0.50	0.5	51	55	0.50	0.0
2	75	0.59	4.9	52	55	0.50	0.0
3	45	0.58	4.9	53	45	0.59	4.7
4	50	0.50	2.7	54	60	0.50	0.5
5	60	0.51	3.6	55	55	0.55	0.0
6	45	0.50	3.9	56	55	0.57	4.0
7	75	0.56	1.7	57	25	0.49	2.0
8	70	0.56	11.2	58	65	0.59	2.1
9	60	0.49	0.0	59	60	0.49	0.0
10	50	0.50	1.6	60	60	0.51	3.6
11	55	0.50	0.0	61	50	0.50	2.7
12	60	0.50	0,0	62	70	0.56	4.4
13	60	0.51	3.6	63	55	0.50	0.0
14	25	0.49	2.0	64	60	0.49	0.0
15	55	0.57	6.6	65	65	0.47	3.1
16	55	0.58	3.4	66	75	0.60	8.0
17	45	0.58	4.0	67	75	0.51	5.2
18	10	0.59	7.6	68	60	0.51	0.0
19	70	0.46	3.0	69	60	0.52	0.0
20	45	0.50	3.9	70	60	0.49	1.6
21	75	0.51	5.2	71	55	0.50	0.0
22	75	0.54	2.9	72	60	0.59	6.6
23	60	0.49	0.0	73	60	0.49	0.0
24	60	0.49	0.0	74	75	0.60	5.6
25	50	0.50	1.7	75	55	0.50	0.0
26	60	0.49	0.0	76	60	0.49	0.0
27	50	0.50	0.8	77	70	0.50	4.5
28	75	0.55	3.2	78	60	0.49	0.0
29	75	0.54	2.9	79	60	0.55	2.1
30	70	0.46	3.0	80	75	0.54	2.9
31	75	0.51	5.2	81	60	0.51	3.6
32	55	0.48	0.0	82	60	0.49	4.4
33	60	0.50	0.0	83	55	0.58	7.1
34	60	0.49	0.0	84	75	0.54	2.9
35	60	0.49	0.0	85	65	0.47	3.1
36	60	0.49	0.0	86	50	0.60	3.1
37	55	0.51	3.2	87	60	0.49	0.0
38	70	0.51	4.5	88	55	0.50	0.0
39	60	0.49	0.0	89	55	0.50	0.0
40	60	0.49	1.9	90	60	0.49	0.0
41	65	0.57	8.5	91	55	0.50	0.0
42	55	0.50	0.0	92	55	0.48	0.7
43	55	0.50	0.0	93	60	0.61	0.1
44	40	0.50	0.0	94	55	0.55	3.0
45	65	0.51	3.2	95	50	0.51	2.7
46	75	0.51	4.9	96	50	0.51	2.7
47	45	0.59	5.7	97	60	0.49	1.9
48	65	0.59	1.6	98	55	0.50	0.0
49	35	0.59	13.4	99	75	0.54	2.9
50	60	0.49	0.0	100	70	0.46	3.0

Table 2. Equal error rates of various online signature verification approaches on DB1

Method	EER (%)
1. Proposed model	**2.61**
2. User Dependent features model [17]	5.06
3. Symbolic Classifier [7]	4.2
4. Linear Programming Description(LPD) [24]	5.6
5. Principal Component Analysis Description(PCAD) [24]	4.2
6. Support Vector Description (SVD) [24]	5.4
7. Nearest Neighbour method Description (NND) [24]	6.3
8. Parzen Window Classifier (PWC) [24]	5.2
9. Mixture of Gaussian Description_3(MOGD_3) [24]	7.3
10. Mixture of Gaussian Description_2 (MOGD_2) [24]	7.0
11. Gaussian Model Description S [24]	4.4

method with other methods where MCYT dataset is used as a validation set. Table-2 shows the performance of our model in comparisons with other models with respect to MCYT-100 dataset. Table-2 shows the minimum EER of different models where 20 genuine signatures are considered for training and 05 genuine and 20 skilled forgeries are considered for testing purpose. From Table-2, it is clear that EER of the proposed model is lowest when compared to that of all other well models for online signature verification. In addition, our model works in lower dimension compared to other models where all 100 global features are used during enrollment.

5 Conclusion

In this paper we propose a novel approach for online signature verification based on recursive subset training. Results clearly indicate that the proposed model is effective in reducing the EER and hence can be used for a real time application. The proposed model is not only results in lowest EER but also works in lower dimension. The main contribution of this paper is the introduction of novel concept of subset training for fixing up writer dependent parameters which is effective in online signature verification and is first in its kind in literature.

Acknowledgement. We would like to express our gratitude to Dr. Julian Fierrez Aguillar, Biometric Research Lab-AVTS, Spain for providing MCYT Online signature dataset. We also thank Deng Cai, Associate Professor, Zhejiang University, China for sharing his work on unsupervised feature selection for multi-cluster data.

References

1. Argones, E.R., Luis, J.A.C.: Online signature verification based on Generative models. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 42, 1231–1241 (2012)
2. Jain, A.K., Griess, F.D., Connel, S.D.: On-line signature verification. Pattern Recognition 35, 2963–2972 (2002)

3. Nanni, L., Lumini, A.: A novel local on-line signature verification system. Pattern Recognition Letters 29, 559–568 (2008)
4. Kashi, R., Hu, J., Nelson, W.L., Turin, W.: A Hidden Markov Model approach to on-line handwritten signature verification. International Journal of Document Analysis and Recognition (IJDAR) 1, 102–109 (1998)
5. Fiérrez-Aguilar, J., Krawczyk, S., Ortega-Garcia, J., Jain, A.K.: Fusion of Local and Regional Approaches for On-Line Signature Verification. In: Li, S.Z., Sun, Z., Tan, T., Pankanti, S., Chollet, G., Zhang, D. (eds.) IWBRS 2005. LNCS, vol. 3781, pp. 188–196. Springer, Heidelberg (2005)
6. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification: the state of the art. Pattern Recognition 2(2), 63–94 (1989)
7. Guru, D.S., Prakash, H.N.: Online signature verification and recognition: An approach based on Symbolic representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), 1059–1073 (2009)
8. Feng, H., Wah, C.C.: Online Signature Verification using a new extreme points warping technique. Pattern Recognition Letters 24(16), 2943–2951 (2003)
9. Faundez, Z., Zanuy, M.F.: On-line signature recognition based on VQ-DTW. Pattern Recognition 40(3), 981–992 (2007)
10. Fierrez, J., Ortega-Garcia, J., Ramos, D., Gonzalez-Rodriguez, J.: HMM-based on-line signature verification: Feature extraction and signature modeling. Pattern Recognition Letters 28(16), 2325–2334 (2007)
11. Kholmatov, A., Yanikoglu, B.: Identity authentication using improved online signature verification method. Pattern Recognition Letters 26, 2400–2408 (2005)
12. Parodi, M., Gomez', J.C., Liwicki, M.: Online Signature Verification Based on Legendre Series Representation. Robustness Assessment of Different Feature Combinations. In: International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), pp. 715–723 (2012)
13. Bajaj, R., Chaudhury, S.: Signature verification using multiple neural classifier. Pattern Recognition Letters 30(1), 1–7 (1997)
14. Baltzakis, H., Papamarkos, N.: A new signature verification technique based on a two stage neural classifier. Engineering Application of Artificial Intelligence 14, 95–103 (2001)
15. Chen, Y., Ding, X.: Online signature verification using direction sequence string matching. In: Proceedings of SPIE, vol. 4875, pp. 744–749 (2002)
16. Wang, K., Wang, Y., Zhang, Z.: On-line Signature Verification Using Segment-to-segment Graph Matching. In: International Conference on Document Analysis and Recognition (ICDAR 2011), pp. 805–808 (2011)
17. Guru, D.S., Manjunatha, K.S., Manjunath, S.: User dependent features in online signature verification. In: Swamy, P.P., Guru, D.S. (eds.) ICMCCA 2012. LNEE, vol. 213, pp. 229–240. Springer, Heidelberg (2013)
18. Wang, J., Wu, L., Kong, J., Li, Y., Zhang, B.: Maximum weight and minimum redundancy: A novel framework for feature subset selection. Pattern Recognition 46, 1616–1627 (2013)
19. Kohavi, R., John, C.H.: Wrapper for feature subset selection. Artificial Intelligence 97, 273–324 (1997)
20. Kira, K., Rendell, L.: A practical approach to feature selection. In: Proceedings of 9th International Workshop on Machine Learning, pp. 249–256 (1992)

21. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
22. Cai, D., Zhang, C., He, X.: Unsupervised Feature Selection for Multi-cluster Data. In: 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010), pp. 333–342 (2010)
23. Guru, D.S., Prakash, H.N., Manjunath, S.: Online Signature Verification: An approach based on Cluster Representation of Global Features. In: Seventh International Conference on Advances in Pattern Recognition, pp. 209–212 (2009)
24. Nanni, L.: Experimental Comparison of One-class Classifier for On-line Signature Verification. Neurocomputing 69, 869–873 (2006)

An Authenticated Transitive-Closure Scheme for Secure Group Communication in MANETS

B. Gopalakrishnan and A. Shanmugam

Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India
bgopal1977@gmail.com, dras_bit@yahoo.com

Abstract. It is essential to provide authentication on mobile nodes in group communication to ensure security and privacy. The nodes that are interested in participating in the group communication form Graphs (V, E). In this paper we authenticate the mobile nodes through transitive closure property of the graph in the routing phase of the On Demand Multicast Routing Protocol (ODMRP) that forms Transitive closure graph. We also performed collaborative group key generation with the nodes defined in the transitive closure graph to accomplish secure communication among the group members. Due to the dynamic nature of nodes in the group, we propose the join and leave algorithm. The rekeying is performed at every change that happens in the (Transitive Closure Graph) TCG. The performance analysis is done by simulation with various protocols with respect to the time taken to joining or leaving the group, time taken for group key generation, rekeying the nodes with respect to renewal of nodes in the group. The proposed system shows our protocol reduces the computational and communicational cost of secure group communication.

Keywords: On demand Multicast Routing Protocol, Group Communications, Authenticated Transitive-Closure Scheme ATCS, Group Key Generation Rekeying.

1 Introduction

A Mobile Ad-hoc NETwork (MANET) is a system of wireless mobile nodes that dynamically self-organize in arbitrary and temporary network topologies. In mobile ad hoc network, nodes can directly communicate with all the other nodes within their radio frequency range; whereas nodes that are not in the direct communication range use intermediate node(s) to communicate with each other. In these two situations, all the nodes that have participated in the communication automatically form a wireless network, therefore this kind of wireless network can be viewed as mobile ad hoc network. These properties make MANET very suitable for group communications.

1.1 Multicast Routing Protocols

Generally there are two types of multicast routing protocols in wireless networks. Tree-based multicast routing protocol. In the tree-based multicasting, structure can be

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 362–369, 2013.
© Springer International Publishing Switzerland 2013

highly unstable in multicast ad-hoc routing protocols, as it needs frequent re-configuration in dynamic networks, an example for these type is Multicast extension for Ad-Hoc On-Demand Distance Vector (MAODV)[1] and Adaptive Demand-Driven Multicast Routing protocol (ADMR)[2]. The second type is mesh-based multicast protocol. Mesh-based multicast routing protocols are more than one path may exist between a source receiver pair, Core-Assisted Mesh Protocol (CAMP) and On-Demand Multicast Routing Protocol (ODMRP)[3] are an example for these type of classification.

This paper is organized as follows Section 2 discuss about many group communication protocols that are developed in the recent years to ensure secure data communication in the group. The Section 3 proposes a new approach to construct transitive closure graph to authenticate the nodes and generate a group key during the route discovery phase of the ODMRP. Due to dynamic nature of the mobile nodes the joining and leaving process is done through the transitive signature of the nodes in the group Section 4 Simulation is performed with our proposed system to establish a secure group communication in MANETs with certain assumptions in the wireless ad hoc networks. It also analyzes various protocols mentioned in related works with our proposed protocol to confirm our protocol benefits over the other protocols.

2 Related Works

Burmester and Desmedt Protocol [4] is an extension of the Diffie-Hellman key distribution system.

$$K_i = (z_{i-1})^{nr}{}_i \,.\, X_i^{\,n-1.}\; X_{i-1}^{\,n-2} \ldots\ldots X_{i-2} \bmod p.$$

That is each group user will come up with the same secret key $k = g^{\,r1r2+r2r3+\ldots+rnrl} \bmod p$, which is the group key shared by all group members. In BD scheme, each group member needs to perform n+1 exponentiations. It also requires a total number of 2n broadcast messages.

Group Diffie–Hellman key exchange [5] is an extension of the DH key agreement protocol that supports group operations. The DH protocol is used for two parties to agree on a common key. In this protocol, instead of two entities, the group may have *n* members. The group agrees on a pair of primes (q and α) and starts calculating in a distributive fashion for the intermediate values. The first member calculates the first value (α_{x1}) and passes it to the next member. Each subsequent member receives the set of intermediary values and raises them using its own secret number generating a new set.

A set generated by the i^{th} member will have *i* intermediate values. For example, the fourth member receives the set:

$$\{\alpha^{x2x3}, \alpha^{x1x3}, \alpha^{x1x2}, \alpha^{x1x2x3}\} \text{ and generates the set}$$
$$\{\alpha^{x2x3x4,}\, \alpha^{x1x3x4,}\, \alpha^{x1x2x4,}\, \alpha^{x1x2x3}, \alpha^{x1x2x3x4}\}$$

The setup time is linear (in terms of n) since all members must contribute in generating the group key. Therefore, the size of the message increases as the sequence reaches the last member and more intermediate values are necessary. With that, the number of exponential operations also increases.

Kim et al. [6] and Perrig [7] use a logical key hierarchy to minimize the number of key held by group members. The difference here is that group members generate the keys in the upper levels using the Diffie–Hellman algorithm rather than using a one-way function. The key of each node is generated from its two children ($k = \alpha^{klkr} mod\ p$). Y. Kim et al. [7] proposed a novel approach to group key agreement by blending binary key trees with Diffie-Hellman key exchange. The resultant protocol suite is very simple, fault-tolerant and secure. We unify the following two important trends in group key management:

1) The use of so-called *key trees* to compute efficiently and update group keys.
2) The use of Diffie-Hellman key exchange hybrids to achieve provably secure and fully distributed protocols.

Harn and Lin [8] proposed an authenticated key transfer protocol based on secret sharing scheme that KGC can broadcast group key information to all group members at once and only authorized group members can recover the group key; but unauthorized users cannot recover the group key. The confidentiality of this transformation is information theoretically secure. Group key generation and distribution, KGC needs to arbitrarily selects a group key and access all public secrets with group members. KGC needs to allocate this group key to all group members in a secure and authenticated method. All communication between KGC and group members are in a broadcast station.

Wei [9] proposed a Hybrid Group Key Management (HGKM) Architecture for Heterogeneous MANET [9]. A heterogeneous MANET forms a two-tier structure, UAV could aid as a trusting center on the ground mobile mainstay nodes and ordinary nodes certification facilities. On the ground floor, each group head node is responsible for the management of a native sub-group of all the ordinary nodes, these nodes can be realized as a common internal node cluster, constitutes the attention on the management of clusters. Cluster head node is usually stronger than ordinary node computing control and constancy, and other anti-attack capability, the general node cluster head node to receive a variety of command and in agreement with the directives for action. All the cluster head nodes establish the first layer of distributed influential agreement.

Chauhan and Tapaswe [10] proposed a secure and efficient Password-Authenticated Group Key Exchange Protocol for Mobile Ad Hoc Networks. This paper shows some security weaknesses in some recently proposed password-authenticated group key exchange protocols. Additionally, a protected and efficient password-authenticated group key exchange protocol in mobile ad hoc networks is future. It only requires continuous round to produce a group session key underneath the dynamic scenario. In other words, the upstairs of key generation is independent of the size of an entire group.

Maheshwari [11] discussed secure key agreement and authentication protocols:

1. This secure key agreement and authentication protocols constructed with distributed collaborative key agreement and authentication protocols for dynamic peer groups. This consists of three interval- based distributed rekeying algorithms, or interval-based algorithms for short, for updating the group key:
   1) The Rebuild algorithm; 2) The Batch algorithm; and 3) The Queue-batch algorithm. The key of node can be generated by BK v = α Kv mod p
2. Where p is any large prime number and α is a primitive root of p.

Kamal [12] proposed a polynomial-based key management scheme for secure intra-group and inter-group communication. He also proposed new approach in group forward and backward secrecy that is a node leaves a group, it can easily compute the new intra-group key based on its old key and the publicly broad-casted data. Similarly, we also show that when a node joins a group, it can discover the old keys.

2.1 An Improved Authenticated Group Key Transfer Protocol Based on Secret Sharing [13]

Confidentiality and authentication are two plain supplies in secure group communication. Specifically, confidentiality safeguards the transmitted message is only familiar for an intended receiver, and authentication guarantees that the communication object is an authorized member. To provide these two basic functions, key establishment protocols are deployed to portion a common one-time session key among group members, which are often classified into key agreement protocols and key transmission protocols. The former includes all members' participation to produce a session key without a trusted third event, but the process of authentication may take an extended time, especially when the amount of members is huge.

3 Proposed System Model

Algorithm for ATCS Protocol

1. **Initialization**

   a. Create a Node with Node Structure in the mesh topology.

Node Structure

N_i	i^{th}Node Identity
Gi	Group ID
α_i	i^{th}Private key
β_i	i^{th}Public key
Path	Array of nodes
Hcount	Hop count
qvalue	Large prime number
Next hop	Next hop node
Status	Source/Intermediate

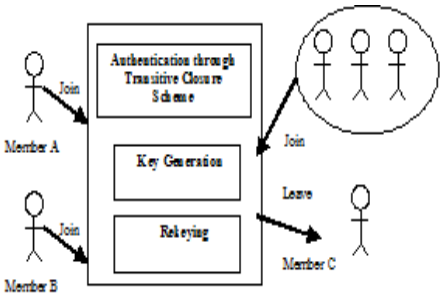


Fig. 1. Model of ATCS

a. Each node in the network agrees with the following parameters:

- Large prime p and q such that q divides p-1
- Two generates g and h of subgroup Gq of order qЄZp* such that the base-g logarithm of h modular p is infeasible for others to compute.
- Let Ni = (N_1, N_2, … , N_n) where n is the number of nodes.

b. Then each node Ni does the following:

Randomly choose two values x_i and y_i from Zp*;

Compute $\alpha_i = x_i \bmod q$ and $\beta_i = y_i \bmod q$;

2. Route Discovery phase

The nodes that are interested in forming a group G will initiate a JoinReq message and floods to all other nodes in the mesh topology.

JoinReq(N_i, α_i , β_i, Hcount, qvalue, Path, Status)

The nodes receiving the **JoinReq**will check for participation in in group communication.

If (Not Participating)

It will just forward the message to its neighboring nodes and save the

Status-Node as Intermediate and add that node to the path

Else

It will send a **JoinAck** message to the Source Node that initiated the **JoinReq** message along the reverse path of the **JoinReq**and Status is set to Source.

JoinAck(N_i, SNid, α_j , β_j, Hcount, Reverse Path, Status);

3. Authentication through Transitive Closure Scheme

Each Source node compute the path verification through transitive closure property

Upon the receipt of $\alpha_j$and $\beta_j$from each neighbor, Source Node *i (SNi).*

Compute

($\alpha ij = \alpha_i - \alpha_j \bmod qvalue$) and
($\beta_{ij} = \beta_i - \beta_j \bmod qvalue$)

The **Source Node** verifies the **Transitive Closure Property**:

$SNi = g^{\alpha}_{i,j} . h^{\beta}_{i,j} \bmod q$

The nodes that satisfies the Transitive closure property will form a group called Transitive Closure Group.

4. Group Key Generation

During the Routing phase of the JoinReq and JoinAck the nodes have exchanged the values of α_i, β_i between the Transitive Closure Group nodes.

Each node 'i' will compute the Group Key GK by

$$GK_i = \sum_{i=1}^{n} N_i . \alpha_i \quad x \quad N_i . \beta_i \bmod p$$

Where 'n' is the Number of Nodes in the TCG.

The Encryption / Decryption can be performed by the group key to have secure communication among the group members.

5. Node Joining the Group

Create a node with Node Structure and floods the JoinReq message

JoinReq(Nid, αi , βi, Hcount, qvalue, Path, Status);

Each node in the TCG will receive the JoinReq message of the new node. Upon receiving the JoinReq all the receiving node will send the JoinAck to the new node.

The New node is added to the TCG and check for Transitive Closure Property. The New node will generate the Group Key using the

$$GK_i = \sum_{i=1}^{n} N_i . \alpha_i \quad x \quad N_i . \beta_i \bmod p$$

The Encryption / Decryption can be performed by the group key to have secure communication among the group members.

6. Node Leaving the TCG

The node wants to leave will send a LReq message

LReq (N_i, G_i, Hcount);

All the nodes in TCG group will send Responce to LRes message to the LReq node as

LRes (N_i,G_i, Path);

The leaving node will check whether all the nodes in the TCG has responded to the LReqthen.

Delete the node from the TCG and Rekeying is performed with the remaining in theTCG.

4 Performance Evaluation

The above protocol (ATCS) is implemented in ns2 simulator. We evaluate the performance of the Transitive closure based algorithm in simulation based experiment. We study their performance in more general setting and also compare the performance of our protocols with other approach specified in the related works.

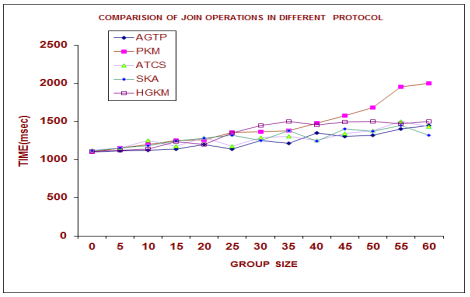


Fig. 2. Comparison of join operations with other protocols

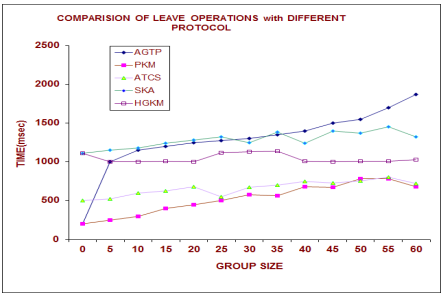


Fig. 3. Comparison of leave operation with other protocols

Figure. 2 shows the performance of protocol ATCS with other protocols. The y axis shows the time taken to generate the group key and x axis shows the number of nodes participated in the group key generation. The time taken is normal i.e. O (n) when the size of the group size increases. Figure. 3 shows the time taken to reconstruct the group key when the node leaves the group is directly proportional to the size of the group members. Figure. 4 shows the time taken to reconstruct the group key when the node leaves the group is directly proportional to the size of the group members. Figure. 5 shows the no. of nodes changed during the authentication of nodes through transitive closure operation, that is, the time to verify the nodes by transitive closure scheme will gradually increase due to increase group size.

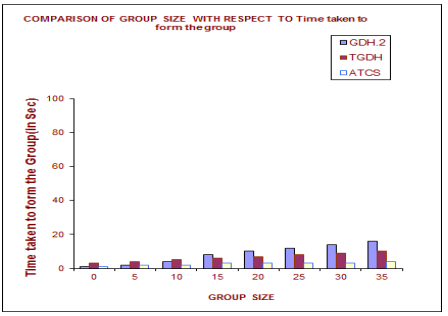


Fig. 4. Comparison with time taken to form the group

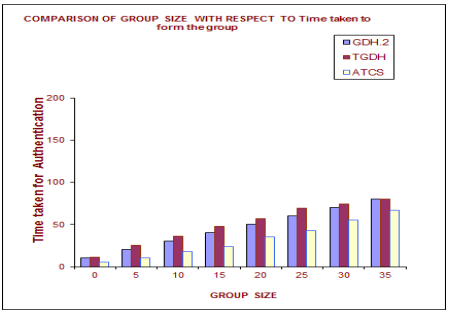


Fig. 5. Comparison with rekeying operation

5 Conclusion

This paper presents a novel scheme ATCS to implement On Demand Multicast Routing Protocol for secure group communication. It uses an efficient way of authenticating the nodes in the route discovery process using transitive Closure scheme. The collaborative group key is generated for secure group communication in MANETs and the rekeying is done due to the mobility of the nodes. The performance of the above protocol is compared with various Authentication protocols in MANETs and Group Key Management protocols in MANETs. This scheme proves to be more suitable for extendable group size, High mobility in the network and secure group communication.

References

1. Royer, E.M., Perkins, C.E.: Multicast Operation of the Ad-hoc On-Demand Distance Vector Routing Protocol. In: Proc. of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), pp. 207–218 (August 1999)
2. Jetcheva, J.G., Johnson, D.B.: Adaptive Demand-Driven Multicast Routing in Multi-Hop Wireless Ad Hoc Networks. In: Proc.of the 2nd ACM International Symposium on Mobile and Ad-hoc Networking & Computing (MobiHOC), pp. 33–44 (October 2001)
3. Lee, S.-J., Gerla, M., Chiang, C.-C.: On-Demand Multicast Routing Protocol. In: Proc. of the Wireless Communications and Networking Conference (WCNC), pp. 1298–1302 (September 1999)
4. Burmester, M., Desmedt, Y.G.: A Secure and Efficient Conference Key Distribution System. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 275–286. Springer, Heidelberg (1995)
5. Steiner, M., Tsudik, G., Waidner, M.: Diffie-Hellman key distribution extended to group Communication. In: SIGSAC Proceedings of the 3rd ACM Conference on Computer and Communications Security, New Delhi, India, pp. 31–37 (1996)
6. Perrig, A.: Efficient collaborative key management protocols for secure autonomous group communication. In: Proceedings of the International Workshop on Cryptographic Techniques and E- Commerce (CrypTEC 1999), Hongkong, pp. 192–202 (July 1999)
7. Perrig, A., Tsudik, G.: Tree-Based Group Key Agreement. ACM Trans. on Information and System Security 7(1), 60–96 (2004)
8. Harn, L., Lin, C.: Authenticated Group Key Transfer Protocol Based on Secret Sharing. IEEE Trans. on Computers 59(6) (June 2010)
9. Wei, C.-Y.: A Hybrid Group Key Management Architecture for Heterogeneous MANET, pp. 565–570. IEEE Computer Society (2010)
10. Chauhan, K.K., Tapaswe, S.: A Secure Key Management System in Group Structured Mobile Ad hoc Networks. IEEE Trans. on Computers, 307–311 (2010)
11. Maheshwari, B.: Secure Key Agreement and Authentication Protocols. International Journal of Computer Science & Engineering Survey (IJCSES) 3(1) (February 2012)
12. Kamal, A.A.: Cryptanalysis of a Polynomial-based Key Management Scheme for Secure Group Communication. International Journal of Network Security 15(1), 59–61 (2013)
13. Liu, Y., Cheng, C., Cao, J., Jiang, T.: An Improved Authenticated Group Key Transfer Protocol Based on Secret Sharing. IEEE Trans. on Computers (2013)

A Novel Ensemble Learning-Based Approach for Click Fraud Detection in Mobile Advertising

Kasun S. Perera[1], Bijay Neupane[1], Mustafa Amir Faisal[2], Zeyar Aung[1,⋆], and Wei Lee Woon[1]

[1] Institute Center for Smart and Sustainable Systems (iSmart), Masdar Institute of Science and Technology, Abu Dhabi, UAE
{kasunsp11,bj21.neupane}@gmail.com, {zaung,wwoon}@masdar.ac.ae
[2] Department of Computer Science, University of Texas at Dallas, USA
mustafa.faisal@utdallas.edu

Abstract. By diverting funds away from legitimate partners (a.k.a publishers), click fraud represents a serious drain on advertising budgets and can seriously harm the viability of the internet advertising market. As such, fraud detection algorithms which can identify fraudulent behavior based on user click patterns are extremely valuable. Based on the BuzzCity dataset, we propose a novel approach for click fraud detection which is based on a set of new features derived from existing attributes. The proposed model is evaluated in terms of the resulting precision, recall and the area under the ROC curve. A final ensemble model based on 6 different learning algorithms proved to be stable with respect to all 3 performance indicators. Our final model shows improved results on training, validation and test datasets, thus demonstrating its generalizability to different datasets.

Keywords: Click fraud, feature extraction, skewed data, ensemble model.

1 Introduction

Smart phones are becoming increasingly popular amongst people of all ages around the world, with new applications and devices being introduced each year. In addition, successive generations of devices possess ever improving internet browsing capabilities, which in turn has resulted in profound changes in the internet usage patterns of large numbers of people, with an increasing proportion of users preferring to access the internet using mobile devices rather than desktop or laptop computers. Similarly, advertising companies around the world have shifted their focus from conventional PCs to mobile computing platforms such as smart phones and tablets. However, in line with this shift, fraudulent activities that were once perpetuated using conventional PCs have now also started appearing amongst the ranks of mobile internet users.

The pay-per-click model that was employed in the conventional internet surfing model has been transferred to the mobile sphere. In the pay-per-click model,

⋆ Corresponding author.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 370–382, 2013.
© Springer International Publishing Switzerland 2013

partners are paid by the amount of internet traffic that they are able to drive to a company's web site. This in turn is measured by the number of clicks received by banners and links associated with their partner accounts. This model gives an incentive for dishonest partners to generate clicks on advertisements on their websites using manual and/or automated techniques. Dishonest advertisers might also generate false clicks on their competitor's advertisements to drain their advertising budgets [1]. These clicks may be generated either manually or through the use of software installed on a mobile phone or a PC.

Though advertising-related fraud has increased greatly over time, perpetuators of fraud still represent only a tiny fraction of the large community of online advertising driven partners. Unfortunately, the activities of this small proportion of fraudulent partners can inflict a very high cost on the profitability of advertisers and other service providers (for example search engines). Thus many research efforts have been undertaken and many approaches have been examined to model the behavior of and subsequently to identify these fraudulent partners. In spite of this, no single model can detect these fraudulent partners with 100% accuracy. In particular, fraud detection in the case of mobile advertising is even harder since many of the features and parameters found on conventional computer systems are not available on mobile phone networks.

Due to the limitations of mobile phone networks, new methods are needed to evaluate the fraudulent behavior of ad partners. New features have to be created using existing parameters to capture the behavior of the partners. Previous studies [2,3] show that these fraud partners try to act rationally when simulating clicking behavior. However, despite (or perhaps due to) these endeavors, their click pattern still deviate from the typical click patterns of legitimate partners.

1.1 Problem Formulation

In this work, we used two tables provided by BuzzCity mobile advertising company (`www.buzzcity.com`) to analyze the behavior of partners and to classify (i.e., to identify) fraudulent partners from legitimate partners. The first table contains the "partner" information as partner id, account no, address and status (in training set only). The second table contains "click" details of all above partners. These details include click id, partner id, ip, agent, category, country, campaign id and referrer. The detailed descriptions of those two tables are available at `www.dnagroup.org/PDF/FDMADataDescription.pdf`.

Since the provided raw data cannot be used directly with any models for classification with sufficient accuracy, we use methods describe in the following sections to add meanings to the tables, aggregate records over partner information and define the behavior of each partner. These formatted data then can be used for classification with methods currently available.

In order to perform the classification of fraud and non-fraud clicks, we propose a novel approach for click fraud detection which is based on a set of new features derived from existing attributes. The proposed model is evaluated in terms of the resulting precision, recall and the area under the ROC curve. A final ensemble

method based on 6 different learning algorithms proved to be stable with respect to all 3 performance indicators.

This paper is an extended version of the competition report [4] to FDMA 2012 [5].

2 Data Preprocessing and Feature Creation

2.1 Preprocessing

Precise and careful data pre-processing and feature creation are critical steps which can greatly improved the overall accuracy of the model. As a first step, we visualized the given data, which include the attributes like iplong, agent, partner id, campaign ID, country, time-at, category, referrer in the clicktable and partner id, address, bank account and label in the partner table. Each attribute in the data was analyzed and evaluated in terms of its effect towards modeling behavior of a partner. Not all of the features would be useful; as such, certain attributes like partner address, bank account from feature creation process were excluded from the modeling process.

After selecting the attributes which were to be considered for further consideration, the partner and click tables from the training set was merged to map the status of clicks to their corresponding partner. Each partner can have one of three possible statuses: "OK" for legitimate partners, "Fraud" for fraudulent ones and "Observation", which indicates that the final status of the partner is unknown and still under scrutiny. To better deal with this, we consider there possible scenarios. In the first scenario, all partners labeled as "Observation" were re-labeled as "OK". In the second, all partners labeled as "Observation" were re-labeled as "Fraud". For the third scenario, all the three labels were retained. Training and testing were performed accordingly on all three scenarios.

2.2 Feature Extraction

Feature extraction is another pre-processing step which can strongly affect the accuracy of the model. Features are numerical indices that quantify the behavior and patterns of every partner. Thus, a properly selected feature should be able to capture properties or trends which are specific to fraudulent partner and be robust towards the evolving patterns of behavior adopted by the said partners.

In order to create features from the selected attributes, we refer to the literature and identify some features that have been used in the Google adsense fraud detection mechanism. Though Google does not reveal the exact features used, the literature provides basic features that can be used in any fraud detection. We also referred to other published research on fraud detection in conventional systems as fraud detection in mobile advertisements is still emerging.

To create features, we process each attribute separately and try to model the relationships of click patterns corresponding a partner by creating a number of parameters (statistical measures) based on that particular attribute. We employ

parameters as maximum, average, skewness, variance etc. We try to create as many features as we can, thus allowing us to capture many characteristics of a partner. In the feature creation phase we were not concerned about the dimensionality of the raw feature because a feature selection procedure would be applied later in the process. Details of the feature creation methods applied to different attributes are as follows.

Attribute - Time-At: Fraudulent partners will often disguise their activities using a variety of tricks such as generating very sparse click sequences, changes in IP addresses, issuing clicks from different computers in different countries and so on. Others stick to the traditional approach of only generating the maximum number of clicks in a given interval. It is important that any prediction system is able to recognize both these attempts at concealment. A number of additional features were derived from the time-at attribute from the given data, with the number of clicks for each partner over different pre-defined time intervals such as 1 min, 5 min, 1 hours, 3hours and 6 hours. The intention was to capture both the short and long term behaviors of the partners. These features were selected as partners often try to act rationally and have constant clicks in very sparse time intervals. We also record the number of clicks each partner receives over different time intervals and then aggregate these values using different measures. We calculated Maximum clicks, Average click, Skewness and Variance in click pattern of each partner for a given time interval. Click Variance for each time interval provide information about click pattern of the partner, whereas Skewness helps to determine the deviation of number of clicks from the average clicks of the partner in a defined time interval. The following figure1 shows all the features we derive from one single attribute "time-at" from given original data.

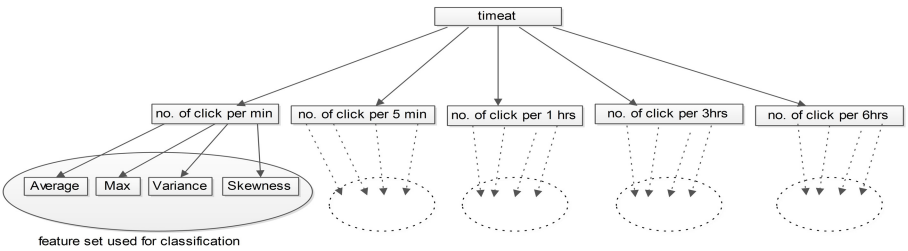


Fig. 1. Feature creation from time-at attribute

Attribute - iplong: IP address is another attribute that can be used to characterize the behavior of a partner, since it is a reflection of the number of computers/mobile devices used or different times at which the user clicks on a particular advertisement. Since many IP addresses are dynamically allocated when users connects via an ISP, it is not unusual for the same user to have different IP addresses. For the given time period of 3 days we observed changes in IP addresses and number of clicks from a given ip address for a given partner id. Then we use

parametric measures over IP address attribute (iplong) to define the behavior of a partner. The following features were created using the iplong attribute.

1. MaxSameIPClicks = for a given partner, count the number of clicks from all different IP addresses he/she have, then obtain the maximum of that
2. NoOfIPs = for a given partner, count the number of clicks from all unique IP addresses he/she have.
3. Click/IP ratio = total clicks for a given partner / total unique ip address associated with that partner.
4. EntropySameIPClicks = for a given partner, count the number of clicks from all different IP addresses he/she have, then obtain the Entropy of that.
5. VarSameIPClicks = for a given partner, count the number of clicks from all different IP addresses he/she have, then obtain the Variance of that.

Some fraudulent partners may try to increase their reward by clicking repeatedly on an advertisement but all of these clicks might come from the same ip. Many clicks originating from the same ip or an unusually large click to ip ratio can be a sign of fraudulent behavior and might place the associated partners under suspicion. Lower variance in the number of clicks from each ip is indicative of a legitimate partner whereas higher variances might indicate a fraudulent partner. Similarly the entropy for the distribution of the number of clicks originating from each IP can be another useful indicator of fraudulent activity.

Attribute - Agent: The Agent attribute is the phone model that user used to browse the web and eventually make clicks on advertisements. As mentioned above, a particular fraudulent user might use one phone, but with many dynamically allocated IP addresses. Thus, we use following measures to derive features from the agent attribute in the set of attributes. The features created using the agent attribute are MaxSameAgentClicks, MaxSameAgentClicks, VarSameAgentClicks, SkewnessSameAgentClicks. These features also calculated using similar method as defined in iplong attribute.

Similarly we define features on Country, Campaign ID as well. When we analyze category with partners, we found that each partner is assigned to only one category, thus we avoid taking category to derive more attributes. But instead we define the prior probability of being fraud for particular category based on the training dataset. We obtained number of users assigned for a given category and then found the number of fraudulent partners in that set to obtain the prior probability of being fraud for a given category. For referrer attribute we derived, referrer/click ratio by obtaining referred clicks over total number of clicks for a given partner. At the end of feature creation process we had 41 different features created from different individual and set of attributes from the dataset. The list of those 41 features is given in `www.dnagroup.org/PDF/FDMAFeatures.pdf`. These carefully crafted features always have threat of over fitting the model, which can be solved by feature selection method.

2.3 Feature Selection

Feature selection is another important step towards building a robust prediction and classification model, and can help to prevent overfitting of the data. We tested our model with different feature selection techniques including Principal Component Analysis (PCA), Common Spatial Patterns (CSP), and wrapper subset evaluation. Of the approaches tested, the wrapper method [6] resulted in the highest accuracy compared to the other methods.

The wrapper method implements a greedy search algorithm for finding the subset of features in space of all feature, where each subset is evaluated in terms of the resulting accuracy when used with the classification model or learning algorithm (in each case, 10-fold cross validation is used to obtain a reliable estimate of accuracy). We also trained and tested our model without performing any feature selection. The results from both approaches are discussed in later sections.

3 Experimental Configuration

3.1 Datasets

The experiment was performed using three distinct sets of data: training, validation, and testing. The training set contains two tables, click-train and partner-train, where click-train includes 8 different attributes and has 3,173,834 instances. Each instance represents a click record for a partner. The table partner-train contains the records for each of the 3,081 legitimate partners along with their labels (fraudulent or legitimate). The validation set also contains similar data with 2,689,005 instances in the click-validation table, while the partner-validation table contains records for 3,064 partners (but without the labels). Various prediction model were built, trained and validated using the training and validation datasets, and one which performs the best on the validation set is selected as the final model. The final model was then evaluated on the test dataset (which remains invisible during the modeling stage). The test dataset also contains two tables: click-test, which contains 2,598,815 instances and partner-test with 2,112 instances.

3.2 Platform and Tools Used

Our experiments were conducted using the Matlab numerical computing environment (`www.mathworks.com/products/matlab`) on a computer with 8GB memory and Intel Core i-7 processor. As the raw click table contains 3 million rows of data, it took a long time to load the data into the Matlab environment. Thus, different approaches were explored to optimize our experimental configuration. We found that SQL is a better choice for creating features from attributes, thus we used MySQL database server (`www.mysql.com`) to load the data and queried every feature we needed from the database. Through a single

script file run on MYSQL server, we were able to generate a CSV file that contained the partners together with 41 features which defined their behavior. This CSV file was then used in WEKA (`www.cs.waikato.ac.nz/ml/weka`) to train, validate and test different models to find the best model. We selected WEKA as it is very user friendly and supports many learning and classification algorithms, and also because of the limited time the we had for our experiments.

3.3 Methods Used

Due to the importance of fraud detection a variety of solutions have been proposed, but none of these solutions work perfectly and produce sufficient accuracy under all conditions. Some solutions work well on the validation dataset but badly on the test dataset. Thus, new methods for fraud detection are always being proposed and evaluated.

Fraud detection is another type of classification which has its own special characteristics. There were many approaches proposed for fraud detection which claim to have higher accuracy. Our approach is to use traditional classification models over data derived from click data. We can tune the parameters in these models to suit with the behavior of our data. It was important that any model used could generalized to work with the training, validation and test datasets and subsequently with any other dataset containing the same features. In order to achieve a stable model, we tried a few different models and over a range of different model parameters; we also selected subsets of features and chose the algorithm with the highest precision, the highest ROC and with a high degree of consistency.

A number of different methods were tried including decision trees, regression trees, artificial neural networks and support vector machines. For each method we also used different learning algorithms, thus each evaluation model is in fact a unique combination of a given classification technique and learning algorithm. After analyzing the results on training and validation data, we found that the decision tree technique was particularly promising and provided very good accuracy. As such, subsequent discussions will cover only decision tree based models.

As mentioned earlier, in any given situation, only a relatively small fraction of the partners are going to be guilty of fraud. As such, any fraud detection data will inevitably be highly skewed (imbalanced), where the number of instances in the "OK" class will greatly outnumber the number of instances in the "Fraud" and "Observation" classes. In our dataset the percentage of Fraud, Observation and OK was 2.336, 2.596 and 95.068 respectively. (The data skewness is one of the fundamental problems faced in many other machine learning applications including network intrusion detection [7], text classification [8], biological data analysis [9], and credit card fraud detection [10], etc.)

The skewed nature of the data forced the prediction model to be biased towards the class with higher population. To deal with the skewed data we used different methods such as resampling and SMOTE. Sampling usually maintains the population of one class and increases (via resampling of the minority class) or decreases (by sub-sampling the majority class) the population of the other

class depending on the type of sampling used. In our experiments we tried both sampling methods followed by randomization of instances. Results obtained both with and without the use of sampling methods are discussed in the results section.

Bagging and boosting are promising ensemble learners that improve the results of any decision tree based learning algorithm. In our research we used Bagging for all the decision tree leaning algorithms. Bagging aggregates multiple hypotheses by the same learning algorithm invoked over different distributions of training data [11]. Bagging helps to generate classifiers with smaller errors on training data as it combines the different hypotheses which individually have a large prediction errors. Metacost which is a form of bagging with cost associated with each training instance to minimize the expected cost was also used as a weak learner with our learning algorithms. We also used random sub space, and logiboost learners with decision tree models.

Based on the results obtained we selected Bagging with decision tree models as the best and most consistent method for classification. A decision tree is a tree like structure, where the classification process starts from a root node and is split on every subsequent step based on the features and their values. The exact structure of a given decision tree is determined by a tree induction algorithm; there are a number of different induction algorithms which are based on different splitting criteria like information gain, gain ratio and the Gini coefficient. As the tree is constructed, over-fitting may occur [12]. Thus, we performed tree pruning to evaluate the performance of the tree and avoid over fitting.

After setting up both the weak algorithms and the main learning algorithm, the resulting models were evaluated using training and validation datasets. After evaluating the results of the different models 6 different models were chosen that obtained the best overall values in precision, recall, and area under ROC curve (AUC). The main purpose of evaluating all three measures is to build a model which can detect a high percentage of fraud cases while maintaining a high degree of precision. The general classification model used is shown in Figure 2.

4 Results

In this section we describe the training and validation results obtained using the approaches mentioned in the previous section. Various decision tree based learning algorithms were analyzed for use as weak learning algorithms in combination with Bagging and Metacost methods. The below are main experiments we performed. Using the algorithms listed in Table 1, different methods were implemented and evaluated to finally reach the most stable model. Those methods are described below. The average precision (AP) metric (Eq. 1) [13] is used to evaluate the different models. The one with the highest AP on the validation set was select as the best model.

$$AP = \frac{1}{m} \sum_{i=1}^{k} Precision(i) \tag{1}$$

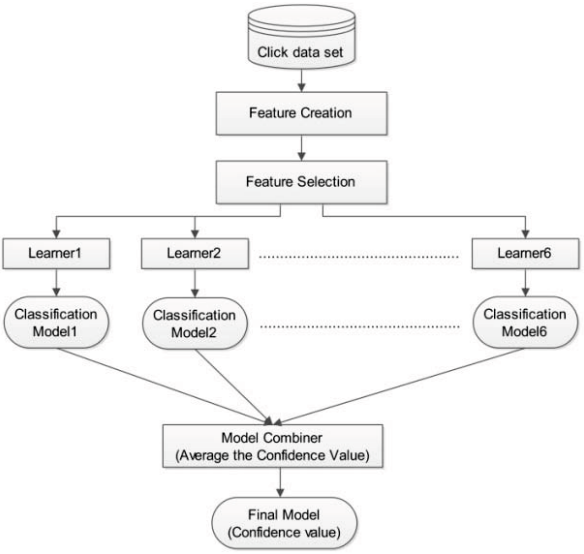


Fig. 2. Classification model

Table 1. Decision tree algorithms and corresponding meta-learning algorithms

Decision Tree Learner	Ensemble Learner
J48	Bagging, Metacost, Logiboost, Random Subspace
Repetition Tree	Bagging, Metacost, Logiboost
Random Forest	Bagging, Metacost, Logiboost

where $Precision(i)$ denotes the precision at cutoff i in the partner list, i.e., the fraction of correct fraud prediction up to the position i, and m is the number of actual fraud partners. If the ith prediction is incorrect, $Precision(i) = 0$ [5].

Results with/without Feature Selection: As mentioned earlier we have generated many features with the training and validation sets provided. To ensure not to make over fitting of the data on the model we have evaluated different algorithm listed above with and without feature selection. We found no significant differences in the results obtained using both approach. The accuracy obtained when using feature selection was actually slightly lower than when no feature selection was used. For example, the AP value obtained using j48 as a learning algorithm without feature selection was 45% whereas with feature selection it was 44.82%. This approach was attempted with all the algorithms and in each case similar results were obtained. One possible reason for this observation is that decision tree algorithms incorporate tree pruning routines which already function as a form of feature selection. As such, inclusion of an additional

feature selection stage will be of limited benefit (or, as was observed in this case, could actually result in a slight drop in accuracy of the resulting classifier).

Results with/without Sampling: As the data was highly skewed, some measures for balancing the data was necessary. We tried to balance the data using two common techniques which were resampling and SMOTE. Resampling and SMOTE performed very well with the training set but performed very badly with the validation set. The results without sampling or SMOTE, in terms of AP, were more than 20% better than the results obtained using sampling and SMOTE.

Results from 2-class/3-class Classifications: Different models were created using the three scenarios as described in Section 2.1. The AP obtained using the J48 tree as the learner and with 2 classes ("Observation" → "OK") was 50.37%, with 2 classes ("Observation" → "Fraud") was 46.1% and with all 3 classes was 45%. Thus we can see that converting all of the "Observation" cases to "OK" and without feature selection as described above was the best approach which gave the highest precision.

Results from Individual Algorithms: We evaluated the prediction performance of different algorithms for all the three . Few algorithms which gave best result alone are mentioned above. There were many algorithms with very low true positive and false negative rates, and which thus were able to obtain very high precision scores. These algorithms were able to obtain high precision because of their low false positive value. We were only interested on algorithms which have high true positive rates and precision. The APs of the different algorithms when applied on the the two class scenario (with "Observation" → "OK") were j48:50.37%, Reptree: 46.82%, and Logiboost: 44.82%.

Results from Combination of Different Learning Algorithms (final approach): With our approach we were able to pass the benchmark score but none of the algorithms alone was able to obtain AP higher than 50.37%. We analyzed the results and found that every algorithm has a drawback, which was either a high false positive rate or a lower rate of true positives. This indicated that choosing any one of the algorithms represented a trade-off between high sensitivity on the one hand, and higher precision on the other. Thus for our next step we combined the results from the different algorithms trained using the 2-class scenario (with "Observation" → "OK") and without feature selection. 6 different algorithms were chosen which obtained higher values for precision, recall and AUC when evaluated alone. This method proved to be the best as we obtained 59.39% AP with the validation set; it also performed well with the final test set, achieving a 46.41% AP. The process of deriving the final click fraud detection model is depicted in Figure 3.

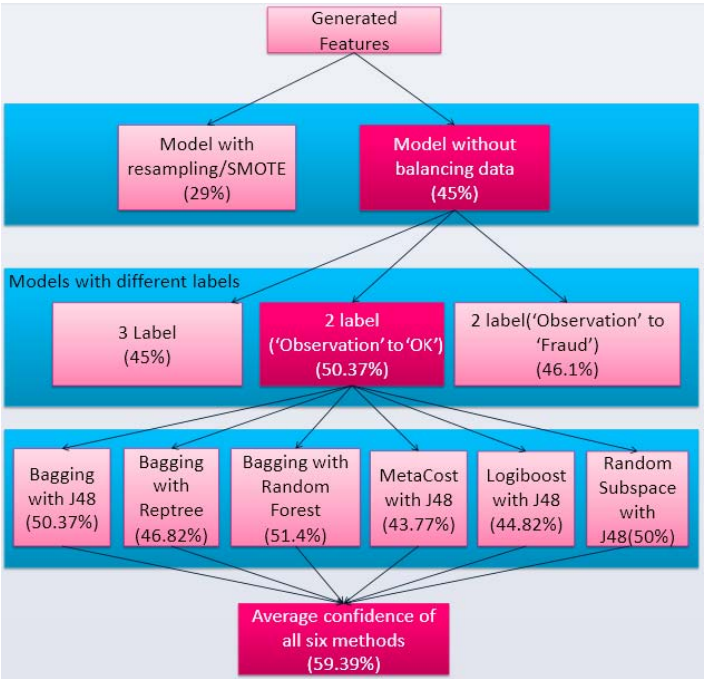


Fig. 3. Process of deriving the final click fraud detection model ensembling 6 different learners. The APs achieved on the validation set are shown in parentheses.

5 Discussions and Conclusion

In this work, we scanned through the features related to fraudulent partners to identify common patterns and to get insights into fraudulent behavior. With the features generated with time-at attribute, we observed that about 90% of the fraudulent partners (∼63) have very small numbers of clicks within the 1min, 5min or 3hr intervals. Also the variance and skewness is very low, which shows that most fraudulent partners operate using small numbers of clicks for the 1 minute – 3 hour time intervals. Though this activity might help them to hide amongst legitimate partners it was observed that most of their clicks sequences were very systematic, which could thus be a sign that these partners as fraudulent. With the agent attribute, we observed very high variance on agents which indicated that many fraudulent partners tried to use large numbers of agents (mobile phone models) to act as a rational user.

Fraud detection in the pay-per-click advertisement model is extremely important as this model is open to abuse by unscrupulous users. Compared to conventional computer systems, fraud detection on mobile phone networks is harder since access to user click information is limited. In this research we perform an experiment to detect fraudulent partners based on click data associated with mobile phone internet surfing. We generate new features based on the

attributes, and use these features to model the behavior of each partner. To achieve a stable prediction model, we perform various feature evaluation techniques, classification algorithms together with different ensemble learners. To generalize the prediction model we combined the results obtained using the 6 learning algorithms which performed best on the training and validation sets. Each algorithms' performance was evaluated based on the average precision score obtained by submitting the results to the competition website. The final results showed that our model performed well with different datasets, and was able to detect a very high number of fraudulent partners.

However, in this work we did not examine feature creation by merging two or more attributes over partners. As future work, we have a plan to identify more features based on the combined attributes which best describe the behavior of a partner. More experiments on feature selection techniques over created features are needed to identify the best features and hence avoid over-fitting.

References

1. Metwally, A., Agrawal, D., El Abbadi, A.: Duplicate detection in click streams. In: Proc. 14th ACM International Conference on World Wide Web (WWW), pp. 12–21 (2005)
2. Kantardzic, M., Walgampaya, C., Wenerstrom, B., Lozitskiy, O., Higgins, S., King, D.: Improving click fraud detection by real time data fusion. In: Proc. 2008 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 69–74 (2008)
3. Li, X., Liu, Y., Zeng, D.: Publisher click fraud in the pay-per-click advertising market: Incentives and consequences. In: Proc. 2011 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 207–209 (2011)
4. Perera, K.S., Neupane, B., Faisal, M.A., Aung, Z., Woon, W.L.: A novel approach based on ensemble learning for fraud detection in mobile advertising. Technical report, International Worshop on Fraud Detection in Mobile Advertising (FDMA) Competition, Singapore (2012)
5. Oentaryo, R.J., et al.: International workshop on fraud detection in mobile advertising (FDMA) competition. In: Conjunction with the 4th Asian Conference on Machine Learning (ACML), Singapore (2012), `http://palanteer.sis.smu.edu.sg/fdma2012`
6. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97, 273–324 (1997)
7. Faisal, M.A., Aung, Z., Williams, J.R., Sanchez, A.: Securing advanced metering infrastructure using intrusion detection system with data stream mining. In: Chau, M., Wang, G.A., Yue, W.T., Chen, H. (eds.) PAISI 2012. LNCS, vol. 7299, pp. 96–111. Springer, Heidelberg (2012)
8. Mladenii, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naive Bayes. In: Proc. 16th International Conference on Machine Learning (ICML), pp. 258–267 (1999)
9. Hugo, W., Song, F., Aung, Z., Ng, S.K., Sung, W.K.: SLiM on Diet: Finding short linear motifs on domain interaction interfaces in Protein Data Bank. Bioinformatics 26, 1036–1042 (2010)

10. Phua, C., Alahakoon, D., Lee, V.: Minority report in fraud detection: Classification of skewed data. ACM SIGKDD Explorations Newsletter 6, 50–59 (2004)
11. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
12. Sahin, Y., Duman, E.: Detecting credit card fraud by decision trees and support vector machines. In: Proc. 2011 International MultiConference of Engineers and Computer Scientists (IMECS), vol. I, pp. 1–6 (2011)
13. Fan, G., Zhu, M.: Detection of rare items with TARGET. Statistics and Its Interface 4, 11–17 (2011)

Neutral Speech to Anger Speech Conversion Using Prosody Modification

Anil Kumar Vuppala[2], J. Limmayya[1], and G. Raghavendra[1]

[1] Department of ECE, RGU IIIT-Nuzvid, A.P, India
[2] International Institute of Information Technology Hyderabad, A.P, India
anil.vuppala@iiit.ac.in, {joga.limmayya,raghava.maths}@gmail.com

Abstract. In this paper, the dynamics of prosodic features are exploited for speech emotion conversion. In particular, emotion conversion of neutral speech to anger speech is accomplished. The database used for analysis of prosody is the Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC). The prosodic features considered for the study are pitch contour, intensity contour, and duration contour. Objective test is performed in terms of average of pitch contour and intensity contour. Subjective listening test results show that the effectiveness of perception of emotion is better in the case of pitch contour modification at the beginning and ending of utterance than for the whole utterance. The results show that the synthesized anger speech is perceived very close to natural anger emotion.

Keywords: Emotion conversion, neutral speech, anger speech, phase vocoder, pitch shift, intensity contour, duration contour.

1 Introduction

The emotion in speech is the extra linguistic information which incorporates expressiveness to tell about mental state of a speaker. Attempts to add emotion effects to synthesized speech or neutral speech have existed for more than a decade and quite a number of smaller studies have been conducted on emotion conversion. Emotions play important role in expressive speech synthesis. Emotional state of a speaker is accompanied by physiological changes affecting respiration, phonation, and articulation. These changes are manifested mainly in prosodic patterns of pitch, energy, and duration. An approach to incorporate emotion into neutral speech is to modify emotion specific parameters to a neutral speech [1]. Therefore, the objective is to analyze and modify these emotion specific parameters of the neutral speech to obtain speech of the target emotion.

Changes in prosodic features from neutral to emotional speech [2] is analyzed and emotion conversion is accomplished. Emotion conversion from neutral to other emotions has interesting applications. For instance, when a computer reads out a story for a child, it will be effective if it is expressed emotionally with corresponding emotions. Besides this there are numerous applications of emotional speech such as announcements in railway stations, employing emotion

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 383–390, 2013.
© Springer International Publishing Switzerland 2013

specific robots in warfare and so on where in expressing emotionally according to circumstances will be much more effective.

It is observed that in the literature there has been lot of work on how the prosodic features vary for different emotions. But there are not much of papers discussing generating emotional speech from neutral utterance. Our interest is that, if we can suitably modify the prosodic parameters of neutral speech, we will be able to synthesize the emotive speech [3]. The motivation has led to development of algorithm to convert neutral prosody to emotional prosody.

The paper is organized as follows. Proposed method for emotion conversion is described in section 4. Section 5 describes Objective test and subjective listening test carried out to estimate the effectiveness of emotion conversion. Finally, Section 6 concludes the paper with a mention on the future scope of the present work.

2 Emotional Database

The database used in this paper, Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC) is recorded using 10 (5 male and 5 female) professional artists from All India Radio (AIR) Vijayawada, India [5]. The total number of utterances in the database is 12000 (15 sentences $*$ 8 emotions $*$ 10 artists $*$ 10 sessions). The eight basic emotions present in the database are: Anger, Sadness, Disgust, Fear, Happiness, Neutral, Sarcastic and Surprise. The speech signal is sampled at 16 kHz, and samples are represented as 16 bit numbers.

3 Prosody Analysis of Anger and Neutral Speech

In this study, we considered 8 utterances of neutral and anger emotions from 6 speakers (3 male and 3 female) for analyzing the prosody variation among various emotions, and also for the comparison of synthesized utterances with natural utterances. The prosodic features of interest are duration patterns, average pitch, pitch contour, and average energy of speech signal [4,2], and average prosody values are tabulated in Table 1. Pitch, duration and strength modification factors obtained by taking ratio of Anger emotion parameters with respect to the neutral emotion are tabulated in 2.These characterize emotion specific information present in speech. From the analysis of the speech corpus and from the literature the several observations are made regarding emotion conversion. Disgust and anger emotions have smaller mean durations, whereas compassion, happy, sarcastic and surprise emotions have comparatively larger mean durations. The extreme emotions like anger, happy, compassion, fear contain emotion specific information in the first words. In comparison to neutral speech, anger is produced with higher and more varied pitch, higher intensity, and shorter duration, and faster attack times at the start of speech [4].

Table 1. Average prosody values for neutral and anger speech

	Duration	Mean pitch	Standard deviation	Energy
	Male			
Anger	1.76	195.60	48.74	203.45
Neutral	1.93	184.37	54.20	160.44
	Female			
Anger	1.80	301.67	80.51	103.36
Neutral	2.04	267.13	77.78	83.42

Table 2. Ratio of features between Anger and Neutral

	F0 mean	F0 Range	Energy	Duration
Neutral	1	1	1	1
Anger	1.15	1.3	1.7	0.84

4 Proposed Method

In the proposed emotion conversion method pitch, intensity and duration of the neutral speech are modified in sequence to generate anger speech. Modification factors presented in Table 2 are used for prosody modification. Techniques for pitch, intensity and duration modifications are presented in following subsections.

4.1 Pitch Contour Modification

The pitch contour which is the characterization of fundamental frequency variation throughout the utterance is tracked and mean value of it is used to discriminate the gender [6]. On the basis of gender, pitch shift factor is determined. The pitch is shifted up using a pitch shifter algorithm function implemented in MATLAB. Figures 1 & 2 show the pitch contour of neutral and anger speech respectively. This function takes a vector of samples in the time domain and shifts the pitch by the number of steps specified. Each step corresponds to a semitone. A phase vocoder is used to time-stretch the signal and then linear interpolation is performed to get the desired pitch shift, and it shown in Figure 3.

4.2 Intensity Contour Raise

Frame by frame analysis is carried out based on the maximum value in a frame. The voiced and unvoiced frames are discriminated and for identified voiced frames the energy is raised. The energy of unvoiced frames is unchanged. The intensity of the whole utterance is raised. For angry speech the energy scale factor is determined from analysis of corpus to be 1.7 with reference to neutral utterance. Figures 4 & 5 show the intensity contour of neutral and anger speech respectively.

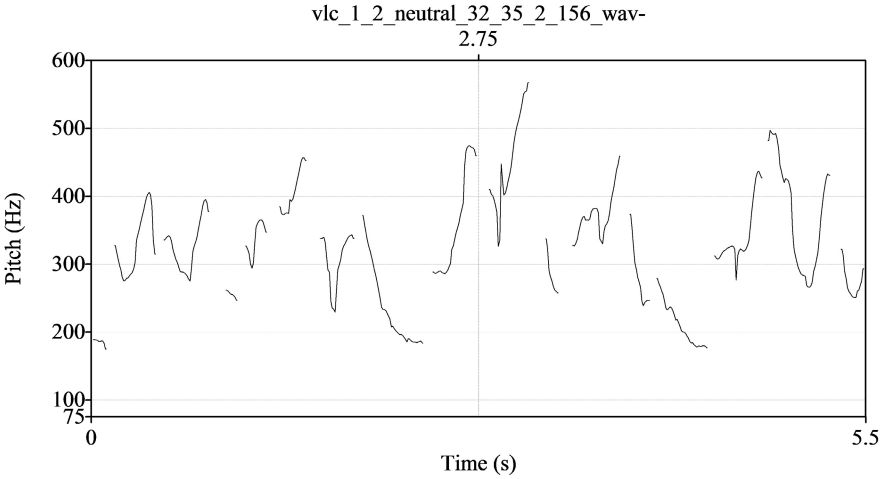


Fig. 1. Neutral pitch contour

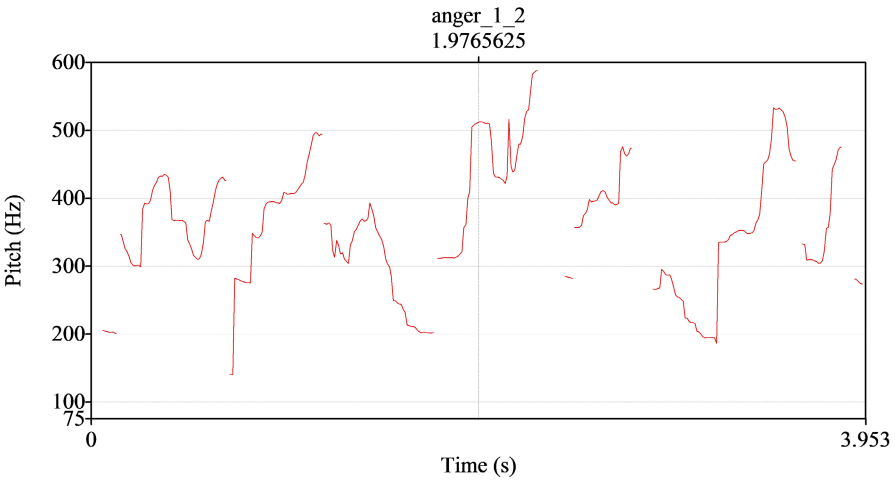


Fig. 2. Anger pitch contour

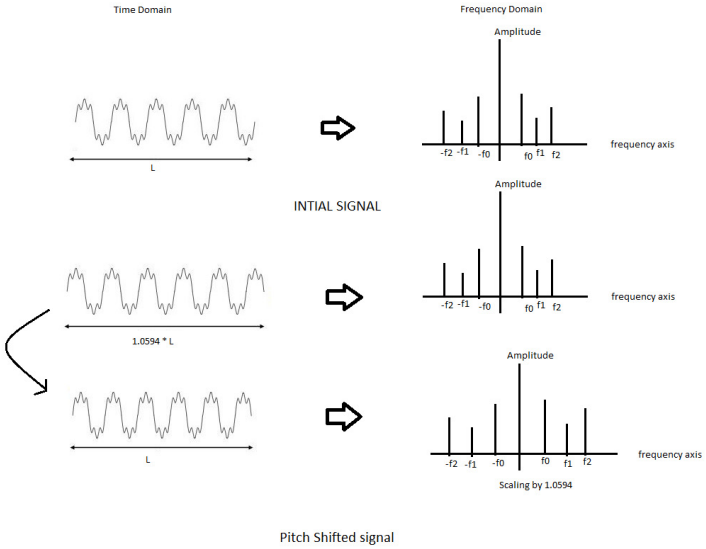


Fig. 3. Pitch shifting by one sample

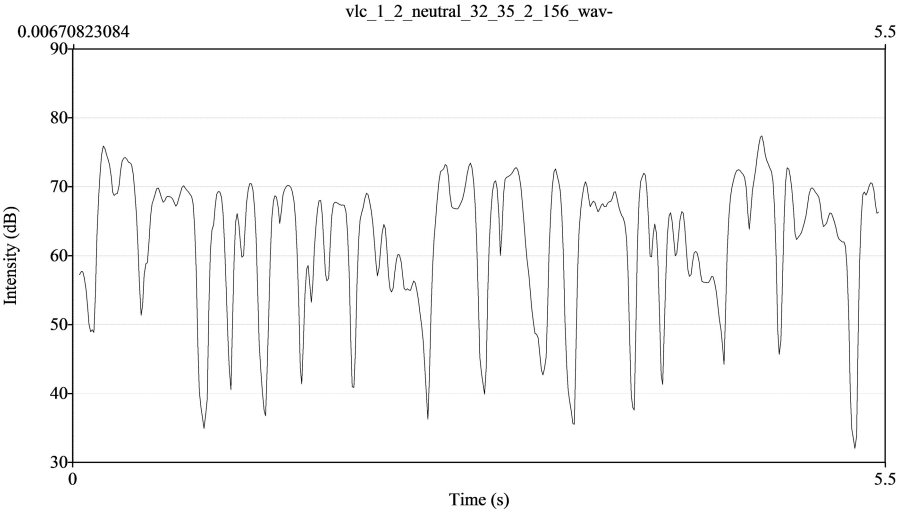


Fig. 4. Neutral intensity contour

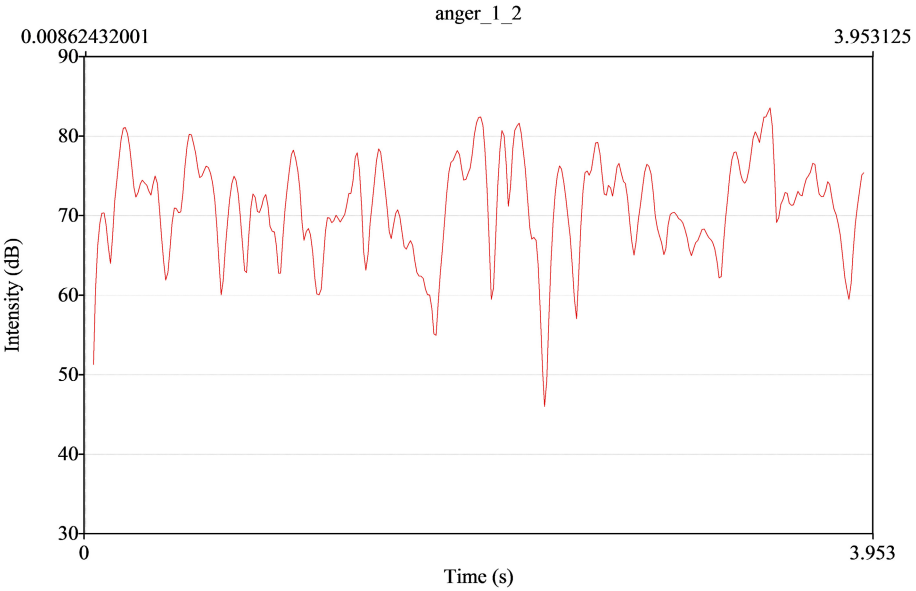


Fig. 5. Anger intensity contour

4.3 Duration Modification

The duration of utterance is modified by suitable scale factor using the phase vocoder MATLAB function. For angry speech the scale factor is determined to be 0.7 compared to neutral. The algorithm calculates the short-time Fourier transform(STFT) of the signal, then builds a modified spectrogram array by sampling the original array at a sequence of fractional time values, interpolating the magnitudes and fixing-up the phases as it goes along. The resulting time-frequency array is inverted back into a sound with inverse STFT. Phase vocoder mainly works on short-time Fourier transform (STFT), typically coded using fast Fourier transform (FFT).

5 Evaluation of Proposed Emotion Conversion Method

Proposed method is implemented in two ways. One is modifying the prosody for whole utterance and second is modifying the prosody only in the beginning and ending of the utterance. Beginning and ending of the utterance is determined by using pause duration. Following subsections describe the objective and subjective testes used for evaluating the proposed method.

5.1 Objective Test

The objective test is performed for the mean values of pitch and energy and they are tabulated in Table 3. The results of objective test show that the mean

Table 3. Prosodic features obtained from neutral and recorded anger speech

	Neutral		Anger		Synthesized Anger	
	Mean Energy	Mean Pitch	Mean Energy	Mean Pitch	Mean Energy	Mean Pitch
Spk 1	72.21	324.37	78.44	406.30	77.32	383.43
Spk 2	73.77	233.96	75.15	316.87	76.15	280.90
Spk 3	70.01	319.90	70.96	327.52	70.08	343.52
Spk 4	76.28	177.87	76.65	206.20	80.59	193.65
Spk 5	68.22	177.02	69.16	213.53	70.33	187.68
Spk 6	73.01	150.35	74.16	176.01	75.17	164.88

values of pitch and energy for recorded anger speech of speech corpus and that of synthesized by the algorithm are very close to one another.

5.2 Subjective Test

The effectiveness of emotion conversion is evaluated by subjective listening test. This evaluation is carried out by 5 human subjects who are working on speech processing in the institute. The human subjects were previously made listened to the actual recordings by professional artists from the database. Subsequently they were given 6 sentences of 3 male and 3 female for giving opinion score on a scale of 5. On the scale 5 represents Excellent and 1 represents Bad perception of the emotion (See Table 4).

The subjective listening test was also performed for pitch change at the beginning and ending of utterance and pitch change for whole utterance. MOS scores

Table 4. Ranking used for judging the quality and perceptual distortion of the speech signal modified by different modification factors

Rating	Speech quality	Level of perceptual distortion
1.	Bad	Very annoying and objectionable
2.	Poor	Annoying but not objectionable
3.	Fair	Perceptible and slightly annoying
4.	Good	Just perceptible but not annoying
5.	Excellent	Imperceptible

Table 5. MOS obtained from different listeners

	Whole Utterance	First & last words
Spk 1	3	3.5
Spk 2	2	2.5
Spk 3	3.5	3
Spk 4	3	3
Spk 5	2.5	3
Spk 6	3	3.5
Average	2.83	3.08

are listed in Table 5. The test results show that the effectiveness of perception of emotion is better in the case of pitch contour modification at the beginning and ending of utterance.

6 Summary and Conclusions

This Paper describes an approach that would modify the emotion of neutral speech signal to produce anger version of it. In this, how prosodic features can be suitably modified to synthesize anger emotion from neutral utterance is investigated. The contours of pitch, energy and duration are exploited for the emotion conversion. It is concluded from the results that modifying pitch contour for beginning words and ending words gives more close to natural anger speech compared to emotion generated by modifying the pitch contour for the whole utterance. The present work can be extended to exploit spectral domain characteristics of emotions for more improved perception of synthesized emotions. The emotion conversion algorithm can be similarly accomplished for other emotions like happy, disgust, sad etc.

References

1. Vroomen, J., Collier, R., Mozziconacci, S.: Duration and intonation in emotional speech. Eurospeech 1, 577–580 (1993)
2. Tao, J., Kang, Y., Li, A.: Prosody conversion from neutral speech to emotional speech. IEEE Transactions on Audio, Speech, and Language Processing 14, 1145–1154 (2006)
3. Rao, K.S., Yegnanarayana, B.: Prosody modification using instants of significant excitation. IEEE Transactions on Audio, Speech and Language Processing 14, 972–980 (2006)
4. Paeschke, A., Sendlmeier, W.F.: Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. In: Speech Emotion, pp. 75–80 (2000)
5. Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Sreenivasa Rao, K.: IITKGP-SESC: Speech database for emotion analysis. In: Ranka, S., et al. (eds.) IC3 2009. CCIS, vol. 40, pp. 485–492. Springer, Heidelberg (2009)
6. Yegnanarayana, B., Murty, K.S.R.: Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Transactions on Audio, Speech and Language Process 17(4), 614–625 (2009)

Expressive Speech Synthesis System Using Unit Selection

Mukta Gahlawat[1], Amita Malik[2], and Poonam Bansal[1]

[1] Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, India
[2] DCRUST, Murthal, Sonepat, India
mukta.gahlawat@gmail.com, amitamalik.cse@dcrustm.org,
pbansal89@yahoo.co.in

Abstract. Speech for realistic environment is hard to achieve. Emotion synthesizing is one way to achieve realistic and natural sounding speech. Use of right emotion in synthesized speech generates the speech which is more effective and natural for listener. The implementation of emotions is very difficult, as word "emotion" has no single definition. There have been various attempts in creating emotional speech synthesis but perfect or near to ideal system has not been developed so far. Our paper is an attempt to create emotional speech synthesizer, where we have used the emotional database recorded in our own voice. We have used unit selection and CART method to implement it. We have taken class room environment for teaching pre-school students with three emotions i.e neutral, happy, sad and tested our synthesizer with twenty listeners and found that listeners have significantly identify the emotional state of speaker.

1 Introduction

Speech technology has been under development for several decades. It provide human communication with machine. This technology is used in number of task like providing aid to the voice-disabled, the hearing-disabled, and the blind. The ability to converse freely with a machine represents the ultimate challenge to our understanding of the production and perception processes involved in human speech communication [1]. Before describing speech synthesis process, we need to understand some basics of speech. This section deals with some basic concepts which are used in speech. Phonetics is the study of human sound speech or equivalent aspects of sign. It deals with the physical properties, physiological production, acoustic properties, auditory perception and neuro physiological status of speech sounds or signs. Oral languages are concerned with three basic areas of study [2]: *Articulatory, Acoustic and Auditory phonetics*. In speech synthesis, we use various acoustic units that are used are described below. *Phone* is the basic speech sound which is not language specific. If phone sound get associated with the language, it is called *Phoneme*. Phonemes are produced by articulation i.e modification of airflow from larynx. These phonemes are combined together to form words. Phonemes are classified into two broad categories viz Vowels and Consonants. *Syllable* is a unit of organization for a sequence of speech sounds. Phonetic sounds are classified under two major groups : vowels and

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 391–401, 2013.
© Springer International Publishing Switzerland 2013

consonants. Prosodic features or suprasegmental features are those aspects of speech which go beyond phonemes. They deal with the auditory qualities of sound the characteristics of the segments i.e individual sounds of speech, e.g place and manner of articulation and voicing for consonants, tongue height, lip rounding, and tenseness for vowels etc.. The main supra segmental features that are used with sound are intonation (variation of pitch), stress (giving emphasis on certain words or phrase in a sentence), pitch (fundamental frequency of vocal cords), tone (use of pitch to distinguish meaning of different words), accent (giving prominence to a particular word, phrase or syllable). There are some additional features along with above mentioned which are responsible for speech variation. These features which make speech of one individual differ from another are : Style variation, voice quality, stress, context and speaking rate.

Speech synthesis is process of conversion of text into speech. Speech Synthesizer is the computer device whose responsibility is to convert input text into equivalent speech signal. Speech Synthesizer or TTS engine consists of two parts i.e front end and back end. Front end contains two modules: text analysis and language analysis module. Backend consists of waveform generator module. Our paper is classified into six sections. Section 2 will discuss research work done on emotional speech synthesis. Section 3 will elaborate design of our emotional speech synthesizer. Section 4 describe the corpus created by us in details. Section 5 shows how our synthesizer is implemented followed by result and conclusion in Section 6.

2 Related Work

Synthesis of robotic speech is not need of system. Now a day's more stress is given on natural sounding speech which can replace a human being. This section will describe work done in field of emotions.

2.1 Emotions

Emotion implementation in speech is fascinating but very difficult research domain.For developing emotional speech synthesizer, it is important to know the meaning of emotion. Emotions can be defined as those mental states with identifiable effects on physiology and behaviour [8]. But this definition is not enough, there are many different perspectives on emotion. The authors in [6] [7] [8] has done research on emotion state and described emotion in four perspective. *The Darwinian perspective:* Charles Darwin's work laid down in his 1872 book *The Expression of Emotion in Man and Animals*, emotions are seen as reaction patterns shaped through evolution. *The Jamesian perspective:* William James in his 1884 article "*What is an emotion*?", the body is seen as essential for the emotion. It states that there is a small but reliable effect of a person's facial expression on his or her subjective emotional experience. *The cognitive perspective:* In cognitive emotion theories, the main

concept is *appraisal*.. Significance of stimulus for individual is determined by appraisal of stimulus and accordingly appropriate response is generated. *The social constructivist perspective:* The "youngest" among the views on what emotions are is attributed by [3] [7]. Here emotions are seen as socially constructed pattern that are learned and culturally shared.

2.2 Emotion Categories

Russell [11] has given Circumplex Model of Affect. Circular ordering of emotion was developed for classification of an emotion. For conceptual similarity are represented by proximity of two emotion categories.

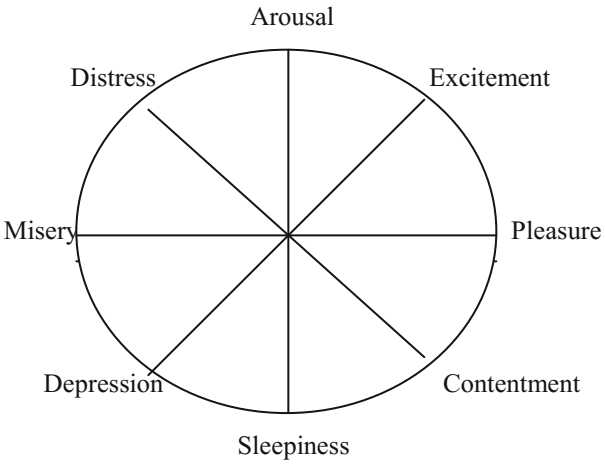


Fig. 1. Circumplex Model of Affect [11]

2.3 Emotions in Speech

For generating the emotion in speech there are some methods in literature. By changing pitch of the input voice, significant changes can occur in the voice and by varying the time elapses we can create emotional sound by giving more space between two words, accent Shape, average Pitch, contour Slope, pitch range, fluent pauses, hesitation pauses, speech rate and stress frequency etc. Quality of voice also determines emotion that more excited voice is loud as compared to a sad voice. Articulation also plays a crucial role in the overall sentence formation to depict different emotions. Various parameters were considered for voice quality and articulation like breathiness, brilliance, loudness, pause discontinuity etc. Acoustic variables are correlated for dimensional description of emotional speech. Many stable correlations have been found [12]. A set of emotional prosody rules were formulated for TTS and provide link between acoustic parameters with 3 emotion dimension i.e. activation, evaluation and power [13]. TTS system with emotion was developed by

[14]. Their work examine the ability to preserve the underlying emotion present in a training corpus. They have considered three emotions “lively,” “sad,” and “angry”. Some of the characteristics of human emotional speech are investigated in [15]. The analysis procedure were done with respect to the perspective of applying the findings to the formant based Text-to-Speech system. Parameters like pitch contour, intensity, speech rate and phoneme duration for four different emotions(anger, fear, joy and grief) were analyzed. When the results of the analysis of the emotional speech data were compared to the corresponding parameters of emotionally neutral human speech, diversion was found. It implies that the modification factor is necessary for the production of emotional synthetic speech. Some of the characteristics of human emotional speech are investigated and the analysis results are reported in [24]. The collection of speech material to be analyzed and the analysis procedure were done. The following parameters have been investigated: pitch contour, intensity, speech rate and phoneme duration for four different emotions: anger, fear, joy and grief. The results were compared to the corresponding parameters of emotionally neutral human speech. They found that when there is diversion from the emotionally neutral parameters values, the modification factor is necessary.

3 Emotional Speech Synthesizer

As described above embedding right kind of emotion in speech leads to generation of speech which sounds very realistic and natural. Various techniques for generation of speech which are available with us in literature. We have taken unit selection algorithm for selection of right unit from emotional corpus. Concatenative technique for speech synthesis is used for joining the units selected from emotional database.

3.1 Software or Tools Used

We have build an emotional database using Audacity [20]. It is a free, easy-to-use and multilingual audio editor and recorder for Windows, Mac OS X, GNU/Linux and other operating systems.. Wavesurfer [21] is used for segment file creation. WaveSurfer is an open source tool for sound visualization and manipulation. For implementation of emotional speech synthesizer we have used TTSBOX [19]. It is a Matlab toolbox for text-to-Speech synthesis. It performs the synthesis of generic English called Genglish. It support an imaginary language obtained by replacing English words by generic words.

3.2 Design of Emotional Speech Synthesizer

Our emotional speech synthesizer is implemented using unit selection algorithm [22], chink and chunk algorithm [23] and CART method. Figure 3 shows the flow chart which clearly describe the design of our speech synthesizer.

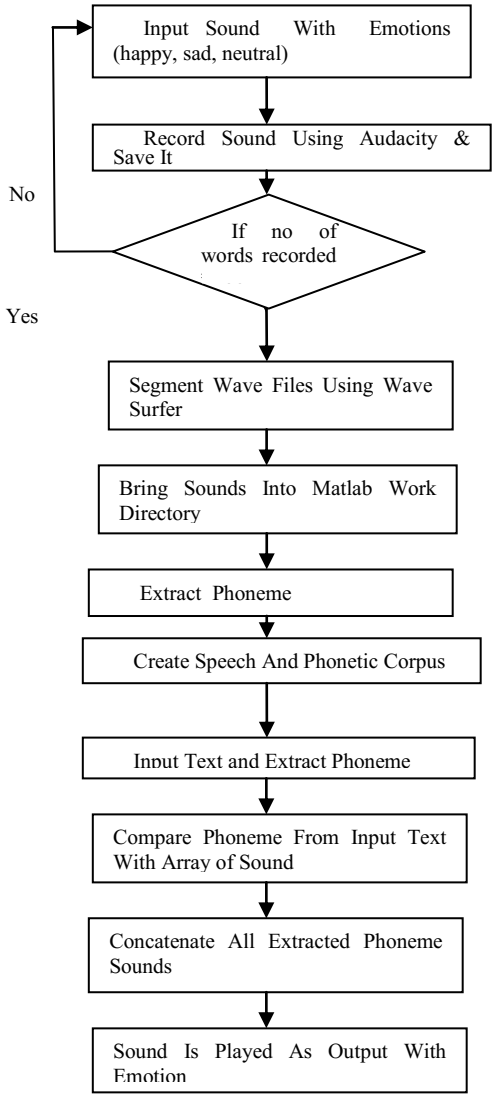


Fig. 2. Flow chart for Emotional Speech Synthesizer

4 Emotional Corpus Creation

This section describes the process of building an emotional database.

4.1 Development of Speech Database Methodology

Speech Corpora design for class room learning environment consists following steps [27].

- Selection of textual content as used in class room.
- Recording of selected textual content.
- Annotation of speech signal.
- Structural storage of corpora.

Table 1. Represents the complete details of database for our Emotional Speech Synthesizer

S No	Environment details for ESS	
1	Environment taken	Nursery class room
2	Approach used	A single voice with a combination of three databases was built for the speaker
3	Algorithm used	Unit selection CART, Chink & chunk algorithm
4	Emotions used	Happy, sad, neutral
5	Speaker Selection	Female voice
6	Tool / Software used	
	1) Recording of database	Audacity
	2) Segmention of file	Wavesurfer
	3) Implementation of synthesizer	Matlab
7	Number of words recorded in each emotion	100
8	Number of utterance of each word	3
9	Total number of words in database	300
10	Type of words	Alphabets , Numbers, Words, Sentences
11	Language used	Indian English

4.2 Speaker Selection

Speaker for a unit selection voice has to be able to speak in the same emotion and same tone of voice constantly for long periods of time. Both these attributes are important for recording an emotional database. We have selected a female speaker. She was asked to read the sentences three times; once in a neutral voice, once in an sad voice, and once in a happy voice. The speaker should have clear voice for better results.

4.3 Which Emotion?

Happy, sad, neutral were taken under consideration for recording emotional database. Emotion "happy" means speaker will speak words or sentences pretending she is in feeling good. "Sad" means she is feeling bad. "Neutral" means she is neither feeling bad nor good.

4.4 Area of Database Recording

A nursery class scene was selected as the particular area of implementation of database. This area was selected so as to depict all the three types of emotions: Happy, Neutral and Sad. This database can be used to generate a system for a Lively Classroom Interaction System. The words were selected accordingly. The words recorded can be arranged to create various sentences. It consisted of 300 words, 100 in each emotion. For example, the following words were included, Hello, Good, Morning, Students, Teacher.

The sentences that can be formed from these words are:

- Hello students.
- Good Morning teacher.
- Good Morning students.

4.5 Recording Procedure

The speech was recorded directly into the computer. The speaker was asked to sit always in the same position and the same distance from the microphone. She was provided with 100 words, which she has to speak in three emotions. We have used audacity for recording purpose. After recording of each data we need to save that data in separate files for each emotion. It was very time consuming procedure as speaker sometime feels very tired of repeating same word in three different emotions. Also by reading again and again sometime speaker commits certain mistakes in reading the data.

4.6 Labeling of Data

After recording of data we have created the database in three emotions. But this database need to be labeled. Without labeling we cannot use our emotional database appropriately. Labeling means first each utterance is associated with labels of

segment, phrases, fundamental frequency, syllables, intonation events and emotion. The segment files for every word in database are created, which shows the time assigned for phonetic transcription. This segment file display the starting time and ending time of each phoneme. These information are used by unit selection algorithm. Segmentation was done with the help of wavesurfer.

Finally our corpus is arranged in alphabetical order which consists of the word, its part-of-speech and the phoneme of each word is written along with it. The database is preprocessed to give a speech corpus used during extraction of phonemes.

5 Implementation

The implementation of emotional speech synthesizer is carried out in a sequence of three major stages, namely, segmentation, association of phoneme, selection of diphone and finally synthesis. The block diagram of the implementation is shown in figure 3.

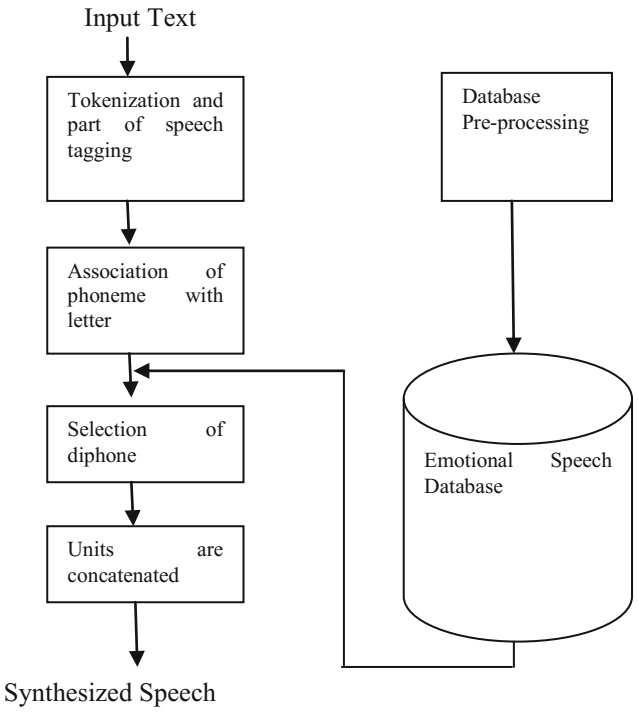


Fig. 3. Block diagram of our System

5.1 Sequence of Steps

First the database is preprocessed. A speech corpus is created where every word in the database is associated with its phoneme and the units are numbered. The user then

inputs the text and the voice from the database is played. The input text is first separated into tokens and tags are associated with each word specifying the part-of-speech of each word. Then phonemes are associated with each letter of the word using the phonetic cart. Then the target sequence if generated and diphones are selected from the database based on the target cost. Then these units are concatenated based on the concatenation cost and the sound is played.

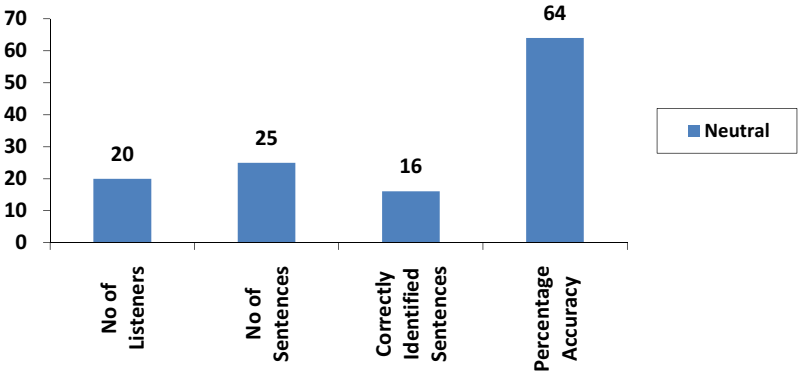


Fig. 4. Result for successful identification of neutral emotion

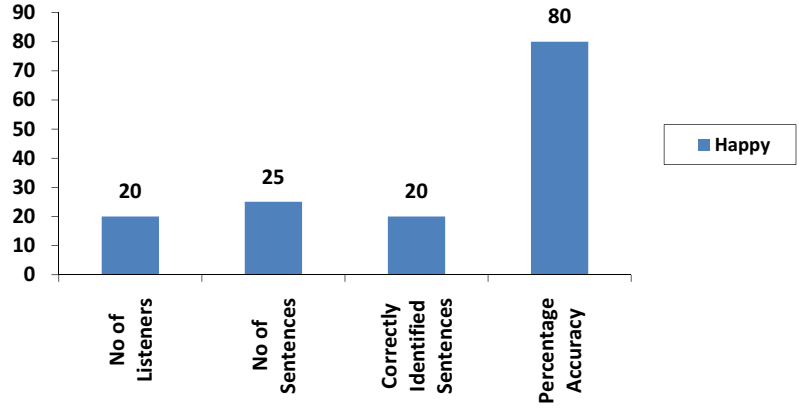


Fig. 5. Result for successful identification of happy emotion

6 Result and Conclusion

The area of work was taken as Nursery school teacher, comprising words related to the specified area only.

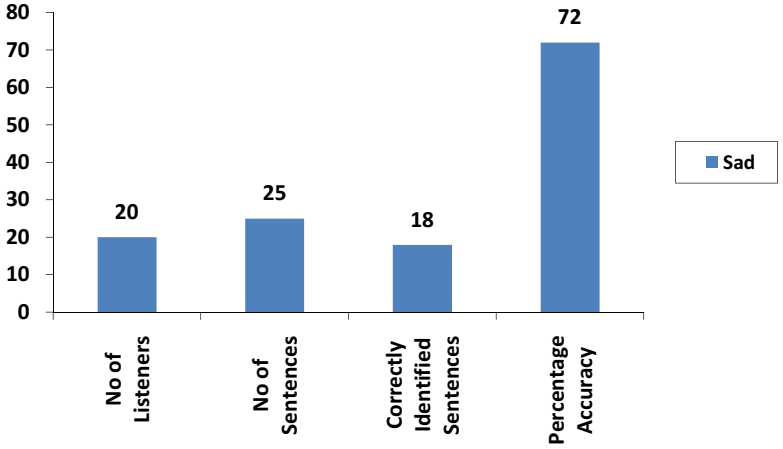


Fig. 6. Result for successful identification of sad emotion

We have tested our emotional speech synthesizer on 20 listeners of various age groups and of different genders. By listening various sentences many listeners have successfully indentified the emotions. Happy emotional sentences have the highest percentage accuracy of 80%, followed by sad 72%. But for neutral emotion, we found the least percentage of correctly identified sentences i.e 64%. Probably the reason is that sad and happy are extreme emotion which can be easily recognizable by listeners but neutral lies in between two extreme, so listeners sometimes were not able to indentify this emotion.

Our future work will be in direction to enhance our Emotional Speech Synthesizer. We can achieve this by enriching our database with more words. Secondly, we will try to add some additional emotions like anger, surprise etc. Also, work can be done in field enhancing the quality of synthesized speech. We can implement this in real life for blind students to teach them and for story telling . Finally if 3-D Text-to-Speech [25] [26] can be added to this, then it will surely give more realistic, natural and very interesting output which can prove to be more useful for visually impaired people.

References

1. Cole, R.A., Zue, V.: Survey of the State of the Art in Human Language Technology, ch. 1, pp. 1–2
2. Jakobson, R.: Structure of Language and Its mathematical Aspects. In: Symposia in Applied Mathematics. AMS Bookstore (1980)
3. Jurafsky, D., Martin, J.H.: Speech and Language Processing, p. 346. Prentice Hall (2008)
4. Roger, W.E.: English Phonemes. Department of English Furman University Greenville, `http://eweb.furman.edu/~wrogers/phonemes/`

5. O'Grady, W., et al.: Contemporary Linguistics: An Introduction, 5th edn. Bedford/St. Martin's (2005)
6. Cornelius, R.R.: Theoretical approaches to emotion. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 3–10 (2000)
7. Schröder, M.: Speech and Emotion Research. An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis. Universität des Saarlandes. Saarbrücken (2003)
8. Hofer, G.O.: Emotional Speech Synthesis. Master of Science Thesis School of Informatics University of Edinburgh (2004)
9. Schere, K.R.: Vocal affect expression: A review and a model for future research. Psychological Bulletin 99, 143–165 (1986)
10. Averill, J.R.: A semantic atlas of emotional concepts. JSAS Catalog of Selected Documents in Psychology 5:330. Ms. No. 421 (1975)
11. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
12. Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S.: Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In: Eurospeech 2001, vol. 1, pp. 87–90 (2001)
13. Schröder, M.: Expressing degree of activation in synthetic speech. IEEE Transactions on Audio, Speech and Language Processing 14(4), 1128–1136 (2006)
14. Eide, E.: Preservation, Identification, And Use Of Emotion. In: A Text-To-Speech System, pp. 127–130. IEEE (2002)
15. Galanis, D., Darsinos, V., Kokkinakis, G.: Investigating Emotional Speech Parameters For Speech Synthesis. In: ICECS 1996, pp. 1227–1230 (1996)
16. Hunt, A.J., Black, A.W.: Unit Selection in a Concatenative Speech Synthesis System Using A Large Speech Database, pp. 373–376. IEEE (1996)
17. Dutoit, T.: An Introduction to Text-to-Speech Synthesis, ch. 6, pp. 150–160. Springer (1997)
18. Timofeev, R.: Classification and Regression Trees (CART) Theory and Applications. Master Thesis, CASE - Center of Applied Statistics and Economics Humboldt University, Berlin (December 20, 2004)
19. TTSBOX available online, `http://tcts.fpms.ac.be/projects/ttsbox`
20. Audacity available online, `http://audacity.sourceforge.net/`
21. Wavesurfer available online, `http://www.speech.kth.se/wavesurfer/`
22. Black, A.W., Campbell, N.: Optimizing selection of Units from Speech Databases for Concatenative Synthesis. In: Proc. of Eurospeech 1995, vol. 1, pp. 581–584 (1995)
23. Liberman, M.Y., Church, K.W.: Text Analysis and Word Pronunciation in Text-to-Speech Synthesis. In: Advances in Speech Signal Processing, New York (1992)
24. Galanis, D., Darsinos, V., Kokkinakis, G.: Investigating Emotional Speech Parameters For Speech Synthesis. In: ICECS 1996, p. 1227 (1996)
25. Moore, et al.:Three Dimensional Speech Synthesis. U.S. Patent 5, pp. 561–736 (October 1,1996)
26. Sodnik, J., Tomažič, S.: Spatial Speaker: 3D Java Text-to-Speech Converter. In: World Congress on Engineering and Computer Science Vol II (WCECS 2009), San Francisco, USA, pp. 1306–1310 (2009)
27. Oliveira, L.C., Paulo, S., Figueira, L., Mendes, C., Nunes, A., Godinho, J.: Methodologies for Designing and Recording Speech Databases for Corpus Based Synthesis. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Marrakech, Morocco (2008)

Edge Based Graph Neural Network to Recognize Semigraph Representation of English Alphabets

R.B. Gnana Jothi and S.M. Meena Rani

V.V. Vanniaperumal College for Women, Virudhunagar, 626001, Tamilnadu, India
gnanajothi_pcs@rediffmail.com, smmeenarani@gmail.com

Abstract. Graph Neural Network based on edges is introduced in this paper and is used to recognize the English uppercase alphabets treating their corresponding graphs as semigraphs. Graph Neural Network(GNN) is a connectionist model comprising of two feedforward neural networks (FNN) called transition network and output network connected by recurrent architecture according to the graph topology. The characteristics of the edges in a graph are considered as input for the transition network and the stabilized output of the transition network are taken as input for the output network. Edge based GNN is trained using error gradient method. Experimental results show that GNN is able to identify all the 26 graphs of alphabets correctly.

Keywords: Graph neural network, Graph structured data, Feedforward network, Recurrent network, Semigraph.

1 Introduction

Graph theory is a subject which originated via puzzles. Euler, the father of graph theory solved the famous Konigsberg bridge problem . Mobius gave the idea of complete graph and bipartite graph and Kuratowski proved that they are planar by means of recreational problems. Kirkman and Hamilton studied cycles on polyhedra and invented the concept called Hamiltonian graph by studying trips that visited certain sites exactly once. Kirchoff's investigation of electric network led to trees in graphs. Supervised neural networks have been developed to deal with structured data encoded as labeled directed cyclic graphs[6]. Binachi et al. have proposed a methodology to process any directed acyclic graph[2]. Micheli[8] has proposed a new approach for learning in structured domains using a constructive neural network for graphs.

Recently GNN has been proposed by Scarseli et al.[16] to process general type of graphs both cyclic and acyclic, directed and undirected. GNN implements a function τ(G,n) ϵ R^m that maps a graph G and one of its nodes n into an m-dimensional Euclidean space. The approximation capabilities of the model are investigated by Scarseli et al.[17] and have shown that under mild generic conditions, many practically useful functions on graphs can be approximated in probability by GNN. Scarselli et al. have applied GNN to different

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 402–412, 2013.
© Springer International Publishing Switzerland 2013

types of problems such as Clique problem[17], Half Hot problem[17], Mutagenesis problem[16], Tree depth problem[17] and subgraph matching problem[16]. Pucci et al.[11] have applied GNN to the movie lens data set and have discussed the problems and limitations encountered by GNN while facing this particular practical problem. Scarselli[15] has modeled GNN for ranking web pages. GNN is applicable for both node focused and graph focused applications of graphical domain. In node focused application, the function τ depends on the node n, so that the classification depends on the properties of the node. In graph focused application, the function τ does not depend on the node n, and the classification is on the graph structured data.

Semigraph introduced by Sampathkumar[14] is a generalization of graph in which every concept of the graph has a natural generalization. Semigraphs are better model than graph in applications where we need to connect several points by an edge[3]. Road networks can be easily modeled by using semigraphs. Traffic routing and traffic density in junctions may be studied through domination in semigraphs[18]. Jeya Bharathi et al.[7] have associated the graph splicing scheme of Freund[5] with semigraphs.

In Pattern recognition, alphabets are recognized using neural networks. Perwej et. al.[9] have represented English alphabet by binary values and recognized using neural networks. Pradeep et al.[10] have used multi layer FNN to recognize offline hand written alphabetical characters using diagonal based feature extraction method for extracting the features of the handwritten alphabets. Dutt et al.[4] have proposed a system which is able to recognize handwritten characters or symbols which has been transformed by scaling or translation or a combination of both and even with noise in them. Using neural network and Euclidean distance metric, Saha et al. [13] have recognized handwritten characters. Reddy et al. [12] have proposed a system capable of recognizing handwritten characters or symbols, inputted by the means of a mouse. Aribowo et al.[1] have built a system which is able to recognize handwritten Latin alphabets in the form of images. The system is developed using hamming network.

In this paper, the English uppercase letters are represented as semigraphs by introducing nodes at places where the sign of the slope of the curve changes and at end points. Edge based GNN is applied to recognize the alphabets treating them as semigraphs. Section 3 describes GNN based on edges, section 4 relates semigraphs and alphabets, section 5 gives the training procedure, in section 6 algorithm and results are discussed.

2 Edge Based Graph Neural Network

A collection of objects having pairwise relation can be represented as graphs. Consider a graph G=(V,E) where V is the set of points called nodes and E is the collection of arcs connecting two nodes of V. Let ne[e] denote the set of edges adjacent to an edge e in E. Let co[e] represents the set of vertices of the edge e at which the neighbouring edges are incident. The label attached to node n and an edge e are given by $l_n \ \varepsilon \ R^c$ and $l_e \ \varepsilon \ R^d$ respectively. A state vector $x_e \ \varepsilon \ R^s$

is attached to each edge e, which represents the characteristics of the edge. The state vector x_e is calculated using FNN which implements a transition function f_w given by

$$x_e = f_w(l_e, l_{co[e]}, x_{ne[e]}, l_{ne[e]}) \tag{1}$$

$$= \sum_{u \varepsilon ne[e]} h_w(l_e, l_{(e,u)}, x_u, l_u) \tag{2}$$

where $l_{(e,u)}\ \varepsilon\ R^k$ represents the characteristics of the node on which the neighbouring edge is incident. For each edge e, h_w is a function of its state and label, label of the neighbouring edge, label of the node at which the edge and its neighbour are incident. Each edge is associated with a feedforward neural network. Number of input patterns of the network depends on its neighbours. h_w is considered to be a linear function. When $h_w(l_e, l_{(e,u)}, x_u, l_u) = A_{n,u}x_u + b_e$ where b_e $\varepsilon\ \mathbf{R}^s$ is defined as the output of feedforward neural network called bias network which implements $\rho_w : \mathbf{R}^d \rightarrow \mathbf{R}^s$; $b_n = \rho_w(l_e)$, $A_{n,u}\varepsilon\mathbf{R}_{s\times s}$ is defined as the output of the feedforward neural network called forcing network which implements $\phi_w : \mathbf{R}^{2*d+k} \rightarrow \mathbf{R}^s$.

$$A_{n,u} = \frac{\mu}{s \times ne[e]} resize(\phi_w(l_e, l_{(e,u)}, l_u))$$

where $\mu\varepsilon(0, 1)$ and resize operator allocates s^2 elements in the output of forcing network to a $s \times s$ matrix. The forcing network and bias network for an edge of the example graph of figure 1 are represented by figure 2 and figure 3 respectively.

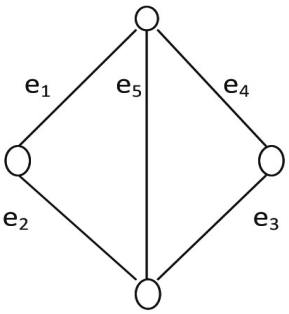


Fig. 1. Example graph

Let x, l denote the vector constructed by stacking all the states and all the labels respectively of the graph. Then equation (1) can be written as

$$x = F_w(x, l) \tag{3}$$

where F_w is the global transition function. Banach fixed point theorem ensures the existence and the uniqueness of solution of equation (1) in the iterative scheme for computing the state

$$x(t + 1) = F_w(x(t), l) \tag{4}$$

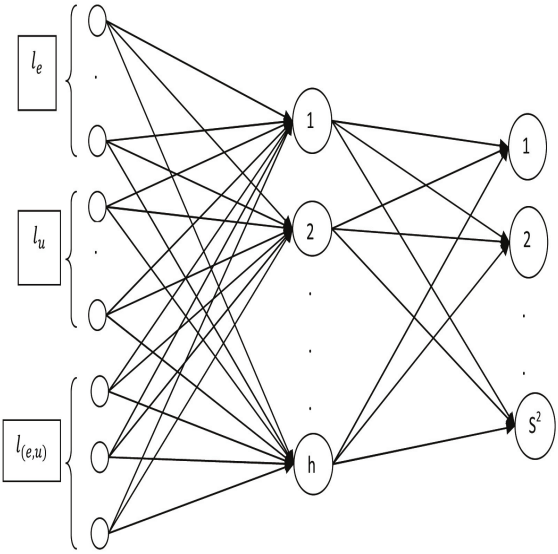


Fig. 2. Forcing network

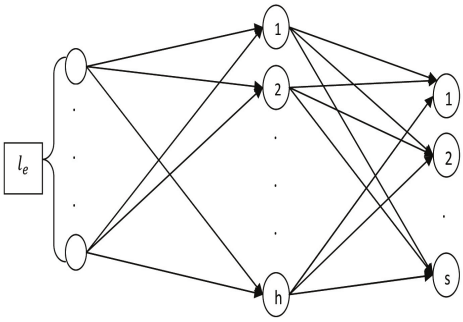


Fig. 3. Bias network

where x(t) denotes the t^{th} iteration of x. Thus the states are computed by iterating

$$x_e(t+1) = f_w(x_e(t), x_{ne[e]}, l_{ne[e]}) \tag{5}$$

This computation is interpreted as a recurrent network that consists of units, namely the transition networks which compute f_w, and the units being connected as per graph topology.

The output of each edge of a graph is produced by a feedforward neural network called output network which takes the stabilized state of the edge generated by the recurrent network and its label as input. For each edge e, the output o_e is computed by the local output function g_w as

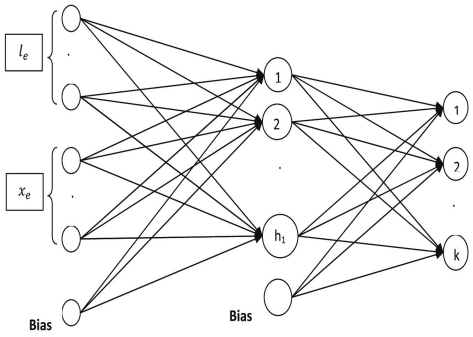


Fig. 4. Output network

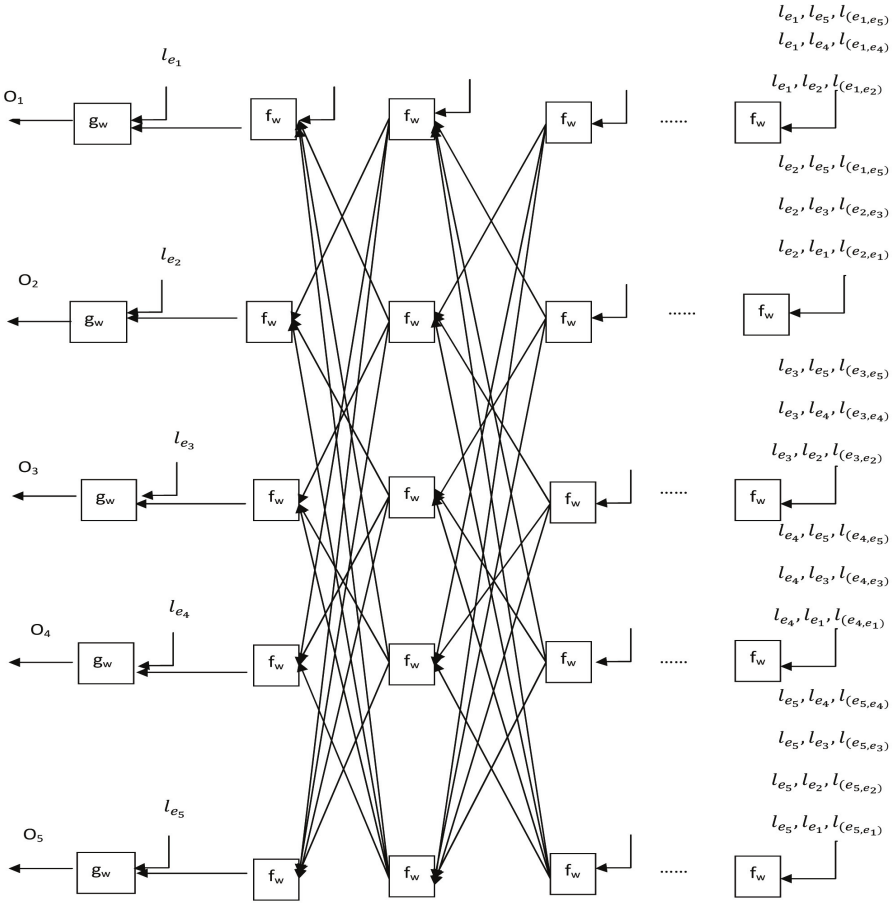


Fig. 5. Graph Neural Network

$$o_e = g_w(x_e, l_e). \tag{6}$$

Global output function is given by $O = G_w(x, l_E)$ where l_E is the vector constructed by stacking all the edge labels. Output network and GNN for the example graph of figure 1 are represented by figure 4 and figure 5 respectively.

3 Semigraph Representation for Alphabets

A semigraph G is a pair (V,X) where V is a non empty set whose elements are called vertices(nodes) of G, and X is a set of n-tuples called edges of G, of distinct vertices for various $n \geq 2$ satisfying the following two conditions.

1. Any two edges have atmost one vertex in common.
2. Two edges $(u_1, u_2, ...u_n)$ and $(v_1, v_2, ...v_m)$ are considered to be equal iff

   $(a) m = n$ and

   (b) either $u_i = v_i$ or $u_i = v_{n-i+1}$ for $1 \leq i \leq n$.

The uppercase English alphabets can be viewed as graphs by introducing nodes at end points, at points where two or more lines with different slopes meet, and at points where the sign of the slope changes in a curve. In the alphabet set A, E, F, H, I, J, K, L, M, N, T, V, W, X, Y, Z are formed by only straight lines. They are represented as graphs by introducing nodes at the end points and at points where two or more lines meet. But alphabets B, C, D, G, O, P, Q, R, S, U are formed with curves and straight lines. They are viewed as graphs by introducing nodes at end points, at point of intersection of lines and curves, and at points where the sign of the slope of the curve changes.

In edge focused GNN, each edge of a graph requires a transition and output network. When graphs are viewed as semigraphs, the number of edges are reduced. In order to reduce the number of transition and output networks in edge focused GNN, the graph of alphabets are treated as semigraphs. The semigraph of alphabets are considered to have edges with two vertices and/or three vertices only for convenience. In the semigraph of alphabets, a straight line with two endpoints is viewed as edge e and the line with three nodes is viewed as edge e'. In the curved portion, as edges of semigraphs can have atmost one node in common, the edges are viewed as e or e' restricting the total number of edges in all the semigraphs to be maximum 4. Fig. 6 gives the semigraph representation of alphabets.

4 Training Procedure

Construct transition and output network for each edge after representing the given problem as a semigraph. The transition network of the edges are to be connected using graph topology. State vectors and weight of the networks of all

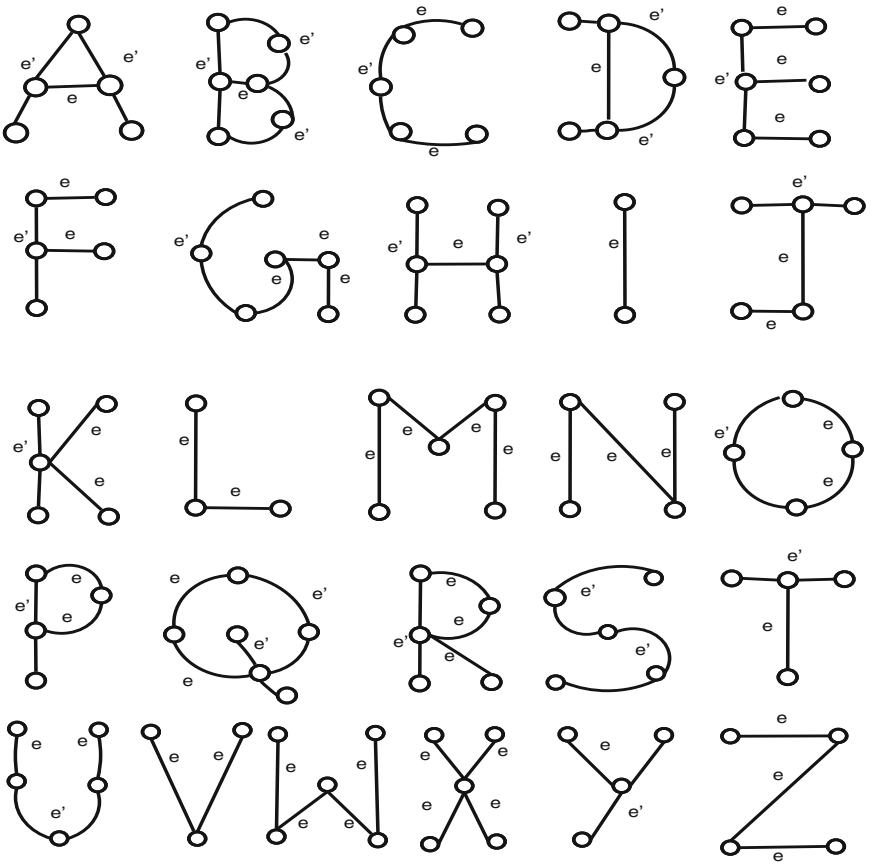


Fig. 6. Semigraph representation of alphabets

edges are to be initialized. The output of each transition network called state vector is computed using feedforward network computations. State vector of each edge is to be computed recursively until it is stabilized. The stabilized state vector and the label of the edge are taken as input for the output network. The error is calculated after calculating the output of each output network. Weights of output networks and then transition networks are to be updated using generalized delta rule of the back propagation algorithm. Based on the error e_w, change of weight for the output network is calculated as

$$\frac{\partial e_w}{\partial w} = \frac{\partial e_w}{\partial O} \frac{\partial G_w(x, l_E)}{\partial w} \tag{7}$$

where $O = G_w(x, l_E)$. Weight updation formula for transition network is

$$\frac{\partial e_w}{\partial w} = b \frac{\partial F_w}{\partial w}(x, l) \tag{8}$$

where $b = \frac{\partial e_w}{\partial O} \frac{\partial G_w(x,l_E)}{\partial x}$. With the updated weights the transition and output network are trained until desired accuracy is obtained.

Algorithm

(1) Find N, the maximum of the number of edges of graphs in the dataset.

(2) Construct N transition and N output networks.

(3) Initialize weights of all the networks.

(4) Select an input pattern (a graph) for GNN.

(5) Initialize state vectors for each edge.

(6) Form input patterns for each transition network.

(7) Find the output of each transition network, using FNN computation, which represents the state vector.

(8) Repeat step 7 recursively until the state vectors are stabilized.

(9) Calculate output for each output network using FNN computation.

(10) Find the error for each output network.

(11) Update the weights of output and transition network using (7) and (8) respectively.

(12) Consider next pattern (graph) for training. Go to step 5.

(13) Compute mean squared error and repeat steps 4 to 12 until desired accuracy is obtained.

5 Results and Discussion

Each alphabet is viewed as semigraph with edges e or e' as specified in section 3 to recognize English alphabets through edge based GNN. The edges are given label of dimension 1. The edges are labeled according to the type of edge, which is a line or curve. If the type of edge is a line, then it may be a horizontal line, vertical line or slant line. The label considered for an edge which is horizontal is 0.01, for a vertical edge it is 0.02, for an edge formed as slant line is 0.03 and for a curve it is 0.04 . In order to make the edge labels distinct, a small noise with mean 0 and standard deviation 1 is added to the labels of the edges so that the semigraphs of the alphabets can be realized by GNN[17].

The semigraph representation of graphs of the alphabets have two types of edges e and e'. Edges of type e has two nodes which are given labels 0.4 and 0.5. Edges of type e' has three vertices which are given labels 0.1, 0.2 and 0.3. The

label $l_{(e,u)}$ of the node of an edge e at which a neighbouring edge u is incident is of dimension 3. The first two components are the labels of the common node in the two edges, and the third component is the angle between e and the incident edge at the common node. The angle charcteristic is represented as -0.18 if the angle is less than 180^0 and 0.18 if it is greater than or equal to 180^0. In Fig. 7, for the semigraph representation of alphabet A, $l_{[e_1,e_2]}$ is $(0.2, 0.4, -0.18)$ and $l_{[e_1,e_3]}$ is (0.1, 0.3, -0.18).

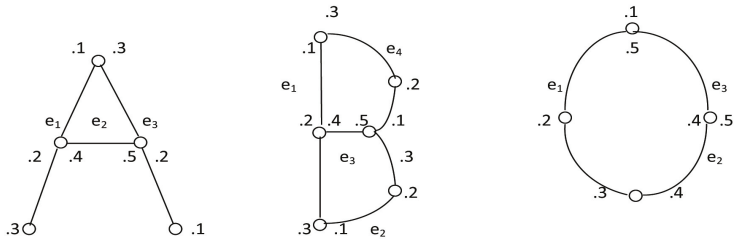


Fig. 7. Semigraph of alphabets A, B,O with labels

In all alphabets, semigraphs are formed with a maximum of four edges. Four transition networks and four output networks are constructed here. The weights of both the networks are initialized with random numbers from $(-1, 1)$. Each edge is adjacent to a maximum of three edges. The dimension of the state of an edge is taken to be 2 and it is initialized with zero vector. Input layer of transition network for an edge e has $2*d+3$ neurons. Both transition and output network have sigmoidal activation function. The number of hidden neurons in the two FNN's are assumed to be different. The number of hidden neurons of the transition networks is considered to be 5. There is no significant change in convergence when the number is changed. The number of hidden neurons in the output network increased as the number of alphabets increased for realization by GNN. The first 20 alphabets are recognized with 16 neurons only. As the graphs of alphabet from 21^{st} alphabet to the 26^{th} alphabet are similar to the graphs of alphabet of first 20 alphabets, the number of hidden neurons in the output network is increased much and it takes 37 hidden neurons for the output network to identify all the 26 alphabets correctly. The experiment is carried out for first 5 alphabets, 10 alphabets, 15 alphabets, 20 alphabets and 26 alphabets for 10 different runs. The results are averaged and represented in Table 1. The learning curve for recognizing all the English alphabets is shown in figure 8. The

Table 1. Time taken to identify the alphabets.

Alphabets	No. of hidden neurons in output network	Time(sec.)	Epoch
First 5	5	23	32
First 10	8	260	377
First 15	12	2209	493
First 20	16	4198	962
All 26	37	5382	688

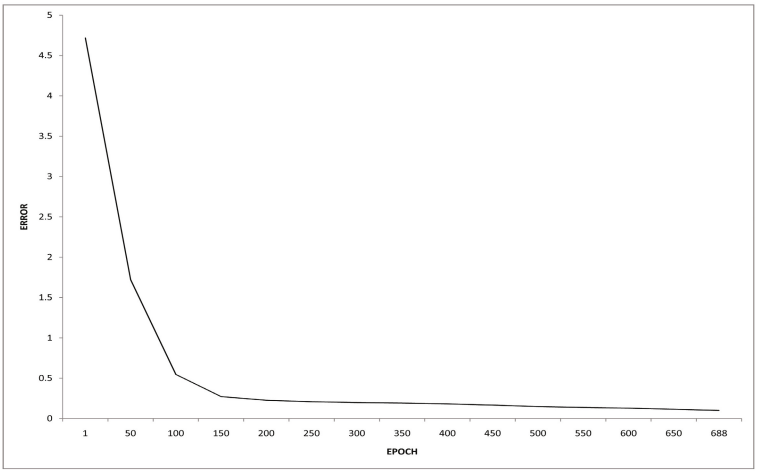


Fig. 8. Learning curve for recognition of all the 26 alphabets

learning rate for both the FNN is 0.01 and momentum up to first 20 alphabets is 0.05 and 0.1 for all the 26 alphabets. The value of μ is considered as 0.001. As the GNN has to identify 26 alphabets, the output for these 26 alphabets are differentiated by considering the binary equivalent of the number that represents the alphabets position in the set of alphabets. Five output neurons are considered in the output network as the binary representation needs 5 binary digits. All edges in a particular alphabet are given the same target as a semigraph is to be identified as a particular alphabet. The termination condition considered in this problem is mean squared error 0.1. The experiment was carried out on a Intel Core 2Quad, 2.49GHZ system.

6 Conclusion

Uppercase English alphabets are represented as graphs and viewed as semigraphs by introducing nodes at points of intersection of lines and curves and at places where the slope of the curve change. Alphabets are drawn as semigraphs having edges with 2 vertices and/or 3 vertices. Edges are labeled based on line or curve which makes the edge. GNN based on edges is used to identify all English alphabets and are identified correctly. Initialization of weights in both transition and output network play an important role in identification. The 21^{st} alphabet to 26^{th} alphabet are similar to some alphabet in the first 20 English alphabets, so more hidden neurons are needed in the output network to recognize all alphabets than first 20 alphabets.

References

1. Aribowo, A., Lukas, S., Handy : Hand written alphabet recognition using hamming network. Seminar Nasional Aplikasi Teknologi Informasi, G-1–G-5 (2007)
2. Bianchini, M., Gori, M., Santi, L., Scarselli, F.: Recursive processing of cyclic graphs. IEEE Trans. Neural Networks 17, 10–17 (2006)
3. Deshpande, C.M., Gaidhani, Y.S.: About adjacency matrix of semigraphs. International Journal of Applied Physics and Mathematics 2, 250–252 (2012)
4. Dutt, V., Dutt, S.: Hand written character recognition using artificial neural network. Advances in Computing 1, 18–23 (2011)
5. Freund, R.: Splicing system of graphs. In: Proceedings of the First International Symposium on Intelligence in Neural and Biological Systems, p. 189. IEEE Computer Society, Washington DC (1995)
6. Hagenbuchner, M., Sperduti, A., Tsoi, A.C.: A self organising map for adaptive processing of structured data. IEEE Trans. Neural Network 14, 491–505 (2003)
7. Jeya Bharathi, S., Padmashree, J., Sinthanai Selvi, S., Thiagarajan, K.: Semigraph structure in DNA splicing system. In: Sixth International Conference on Bio-inspired Computing- Theories and Applications. IEEE Conf. publication 27-29 (2011)
8. Micheli, A.: Neural network for graphs: A contextual constructive approach. IEEE Trans. on Neural Networks 20, 498–511 (2009)
9. Perwej, Y., Chaturvedi, A.: Neural networks for hand written English alphabet recognition. International Journal of Computer Applications 20, 1–5 (2011)
10. Pradeep, J., Srinivasan, E., Himavathi, S.: Diagonal based feature extraction for hand written alphabet recognition system using neural network. In: IEEE Explore Digital Library, International Conference on Electronics Computer Technology, vol. 4, pp. 364–368 (2011)
11. Pucci, A., Gori, M., Hagenbuchner, M., Scarselli, F., Tsoi, A.C.: Applications of Graph neural networks to Large-Scale Recommender Systems some results. In: Proceedings of International Multiconference on Computer Science and Information Technology, pp. 189–195 (2006)
12. Reddy, S.K.D., Rao, S.A.: Hand written character recognition using back propagation network. Journal of Theoretical and applied Information Technology 5, 257–269 (2005)
13. Saha, S., Som, T.: Hand written character recognition by using neural network and Euclidean distance metric. International Journal of Computer Science and Intelligent Computing 2, 1–5 (2010)
14. Sampathkumar, E.: Semigraphs and their applications. Report on the DST(Department of Science and Technology) project submitted to DST, India (May 2000)
15. Scarselli, F., Yong, S.L., Gori, M., Hagenbuchner, M., Tsoi, A.C., Maggini, M.: Graph neural networks for ranking web pages. In: Proceedings of the 2005 IEEE/WIC/ACM Conference on Web Intelligence, Washington, DC, USA, pp. 666–672 (2005)
16. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. on Neural Networks 20, 61–80 (2009)
17. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: Computational capabilities of graph neural networks. IEEE Trans. on Neural Networks 20, 81–102 (2009)
18. Venkatakrishnan, Y.B., Swaminathan, V.: Bipartite theory of semigraphs. WSEAS Trans. on Mathematics 11, 1–9 (2012)

Neural Rotor Time Constant Estimation for Indirect Vector Controlled Induction Motor Drives

Moulay Rachid Douiri, Ouissam Belghazi, and Mohamed Cherkaoui

Mohammadia Engineering School, Department of Electrical Engineering,
Ibn Sina Avenue, 765, Agdal-Rabat, Morocco
douirirachid@hotmail.com

Abstract. In this paper a back-propagation neural networks is employed for indirect field oriented control drive systems to identify the rotor time constant using measurements of the stator voltages and currents and the rotor speed of the induction motor. The neural network model outputs are compared to the desired values, and the total error between the desired and the estimated state variable is then back-propagated to adjust the weights (rotor time constant) of the neural model, so that the output of this model will coincide with the desired value. The back-propagation mechanism is easy to derive and the estimated rotor time constant tracks precisely the actual motor rotor time constant. The theoretical analysis as well as the simulation results to verify the effectiveness of the new approach is described in this paper.

Keywords: artificial neural networks, indirect field oriented control, induction motor drives, rotor time constant.

1 Introduction

In the indirect method of vector control, the unit vectors are generated by using the measured rotor speed and the calculated slip frequency. The calculation of slip speed depends on the rotor time constant which changes significantly with temperature. An error in the slip speed calculation produces an error in the unit vectors, which results in coupling between the flux and torque-producing currents due to axis misalignment [1],[2]. This results in a sluggish torque response with possible overshoot or undershoot and a steady-state error. There is, therefore, a requirement to track the changes in the rotor time constant with a parameter identifier.

Numerous methods [3],[4],[5] have been proposed to circumvent this problem. However, all of these methods identify the rotor time constant by injecting test signals, which disturb the normal operating functions of the drive, into the motor. Some more detailed sensitivity investigations of EKF applications are published in [6] and [7]. An extended version of the reactive power method is examined in [8]; the influence of load and stator frequency on the parameter sensitivity is not considered.

Some authors use intelligent methods [9],[10] to estimate the rotor time constant. Ba-Razzouk, A., et all. [11] proposed a rotor time constant estimation method using

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 413–422, 2013.
© Springer International Publishing Switzerland 2013

the back-propagation neural network from motor terminal voltage, current and frequency, formed by log-sigmoidal neurons, and comprises 5 inputs, 6 neurons on the first hidden layer, 6 neurons on the second and one output neuron. It converges to a sum squared error of $9.26.10^{-4}$ after 14,500 iterations (with randomly initialized weights and biases in the beginning of the training).

The main objective of this research is to estimate the rotor time constant of an induction motor drive that provides more precise results than obtained by Ba-RazzoukA., et al. [11]. This paper is organized as follows: in section 2, describes principles of indirect vector control. In section 3, we present the theoretical analysis of the parameter sensitivity. In section 3, we develop the steps for training a neural rotor time constant. Section 4 presents the simulation results obtained for this approach. Paper ends with a brief conclusion in Section 5.

Nomenclature

R_s , R_r	Stator and rotor resistances [Ω]
i_{ds} , i_{qs}	Direct and quadrature stator currents [A]
i_{dr} , i_{qr}	Direct and quadrature rotor currents [A]
v_{ds} , v_{qs}	Direct and quadrature stator voltages [V]
v_{dr} , v_{qr}	Direct and quadrature rotor voltages [V]
L_s , L_r, L_m	Stator, rotor and mutual inductance [H]
λ_{ds}, λ_{qs}	Direct and quadrature stator fluxes [Wb]
λ_{dr}, λ_{qr}	Direct and quadrature rotor fluxes [Wb]
T_{em}	Electromagnetic torque [N.m]
ω_r, ω_e, ω_{sl}	Rotor, synchronous and slip frequency [rad/s]
τ_r	Rotor time constant (L_r/R_r) [s]
J	Inertia moment [Kg.m^2]
n_p	Number of pole pairs
$\sigma = 1 - \frac{L_m^2}{L_s L_r}$	Leakage coefficient

2 Indirect Field Oriented Control

The dynamic behavior of a three-phase and three-wire induction machine with a squirrel-cage rotor can be described in a *d-q* rotation reference frame. With proper constraints, the d-q frame can be made to rotate synchronously with the stator or rotor flux. The so-called field oriented control allows the rotor flux to be perfectly aligned with the d axis of the *d-q* frame. It follows that:

$$\lambda_{qr} = 0,\ \dot{\lambda}_{qr} = 0 \tag{1}$$

Moreover, the electromagnetic torque equation can be expressed in terms of the stator current and rotor flux linkage as:

$$T_{em} = \frac{3}{2} n_p \frac{L_m}{L_r} \left(\lambda_{dr} i_{qs} - \lambda_{qr} i_{ds} \right) \tag{2}$$

Moreover, the synchronous angular velocity (ω_e) in the indirect field-oriented mechanism is generated by using the measured rotor angular velocity (ω_r) and the following estimated slip angular velocity:

$$\omega_{sl}^{*} = \frac{i_{qs}^{*}}{\tau_r i_{ds}^{*}} \tag{3}$$

where asterisk (*) is the signal command.

Therefore, the value of rotor circuit parameters is necessary for calculating slip frequency command. However, these parameters, especially the value of rotor resistance, vary with the operating temperature as well as saturation level strongly. As a result, the performance of torque control is deteriorated significantly both in steady state and transient condition.

3 Theoretical Analysis of the Rotor Time Constant Sensitivity

The aim of this Section is to derive an expression:

$$\frac{\hat{\tau}_r - \tau_r}{\tau_r} = \frac{\Delta \tau_r}{\tau_r} \tag{4}$$

One gets for the *d*-axis voltage:

$$\left. \frac{\Delta \tau_r}{\tau_r} \right|_{\Delta R_s = 0} = \frac{\dfrac{\Delta(\sigma L_s)}{\sigma L_s} \left(1 + \dfrac{i_{sq}^2}{i_{sd}^2} \right)}{\dfrac{1-\sigma}{\sigma} - \dfrac{\Delta(\sigma L_s)}{\sigma L_s} \left(1 + \dfrac{i_{sq}^2}{i_{sd}^2} \right)} = \frac{\dfrac{\Delta(\sigma L_s)}{\sigma L_s}}{\dfrac{i_{sq}^2}{i_s^2} \dfrac{1-\sigma}{\sigma} - \dfrac{\Delta(\sigma L_s)}{\sigma L_s}} \tag{5}$$

$$\left. \frac{\Delta \tau_r}{\tau_r} \right|_{\Delta(\sigma L_s) = 0} = \frac{-\dfrac{\Delta R_s}{R_s} \left(1 + \dfrac{i_{sq}^2}{i_{sd}^2} \right)}{\dfrac{i_{sq}}{i_{sd}} \omega_s \dfrac{L_s - \sigma L_s}{R_s}} = \frac{-\dfrac{\Delta R_s}{R_s}}{\dfrac{i_{sd} i_{sq}}{i_s^2} \omega_s (1-\sigma) t_s} \tag{6}$$

For the *q*-axis voltage, the following results are obtained:

$$\left.\frac{\Delta\tau_r}{\tau_r}\right|_{\Delta R_s=0} = \frac{\frac{\Delta L_s}{L_s}\left(1+\frac{i_{sq}^2}{i_{sd}^2}\right)}{(1-\sigma)\frac{i_{sq}^2}{i_{sd}^2}} = \frac{\frac{\Delta L_s}{L_s}}{(1-\sigma)\frac{i_{sq}^2}{i_s^2}} \tag{7}$$

$$\left.\frac{\Delta\tau_r}{\tau_r}\right|_{\Delta L_s=0} = \frac{\frac{\Delta R_s}{R_s}\left(1+\frac{i_{sq}^2}{i_{sd}^2}\right)}{\frac{i_{sq}}{i_{sd}}\omega_s\frac{L_s-\sigma L_s}{R_s}-\frac{\Delta R_s}{R_s}\left(1+\frac{i_{sq}^2}{i_{sd}^2}\right)} = \frac{\frac{\Delta R_s}{R_s}}{\frac{i_{sd}i_{sq}}{i_s^2}\omega_s(1-\sigma)t_s-\frac{\Delta R_s}{R_s}} \tag{8}$$

The power balance methods yield quadratic equations for $\Delta\tau_r/\tau_r$ with results being essentially less clearly arranged and for that reason not given here.

The following conclusions can be derived from (5), (6), (7) and (8):

- All equations contain the expression i_{sq}/i_{sd}, which might be interpreted as load factor, indicating that the τ_r detuning more (5) or less (7) depends on load torque and flux level.
- The functions $\Delta\tau_r(\Delta R_s)$ will almost vanish at higher frequencies, whereas for $\Delta\tau_r(\Delta L_s)$ and $\Delta\tau_r(\Delta(\sigma L_s))$, respectively, no stator frequency dependency must be expected.
- The magnitude of the τ_r detuning depends on the machine parameters in different ways and thereby on the special machine under investigation.

4 Neural Rotor Time Constant Estimator

A rotor flux observer of induction machine can be represented by the following equation.

$$\begin{bmatrix}\frac{\dot{\lambda}_{dr}}{dt}\\ \frac{\dot{\lambda}_{qr}}{dt}\end{bmatrix} = \begin{bmatrix}-\frac{1}{\tau_r} & -\omega_r\\ -\omega_r & -\frac{1}{\tau_r}\end{bmatrix}\begin{bmatrix}\lambda_{dr}\\ \lambda_{qr}\end{bmatrix} + \frac{1}{\tau_r}\begin{bmatrix}i_{ds}\\ i_{qs}\end{bmatrix} \tag{9}$$

We can write (9) as:

$$\dot{\lambda}_{dr} = A\lambda_{dr} + Bi_{dr} + \omega_r\lambda_{dr} \tag{10}$$

$$\dot{\lambda}_{qr} = A\lambda_{qr} + Bi_{qr} + \omega_r\lambda_{qr} \tag{11}$$

Then an error function can be defined as:

$$\varepsilon(\xi) = \dot{\lambda}_{dr} - A\lambda_{dr} - Bi_{dr} - \omega_r \lambda_{dr} \tag{12}$$

$$\varepsilon(\xi) = \dot{\lambda}_{dr} - A\lambda_{dr} - Bi_{dr} - \omega_r \lambda_{dr} \tag{13}$$

where $\xi = [a_{11}, a_{12}, a_{21}, a_{22}, b_{11}, b_{12}, b_{21}, b_{22}]$. Let $\int f$ denotes $\int \left(\frac{1}{t_f}\right) f(t)dt$, where t_f is the is the total integration time, then we can define a cost function as:

$$\mathrm{E}(\xi) = \frac{1}{2}\int \varepsilon^t(\xi)\varepsilon(\xi) \tag{14}$$

Hence,

$$\mathrm{E}(\xi) = \frac{1}{2}\int \left(\dot{\lambda}_{dr} - A\lambda_{dr} - Bi_{dr} - \omega_r \lambda_{dr}\right)^t \left(\dot{\lambda}_{dr} - A\lambda_{dr} - Bi_{dr} - \omega_r \lambda_{dr}\right) \tag{15}$$

$$\mathrm{E}(\xi) = \frac{1}{2}\int \left(\dot{\lambda}_{qr} - A\lambda_{qr} - Bi_{qr} - \omega_r \lambda_{qr}\right)^t \left(\dot{\lambda}_{qr} - A\lambda_{qr} - Bi_{qr} - \omega_r \lambda_{qr}\right) \tag{16}$$

From the optimization theory [12] we obtain:

$$\frac{d\xi}{dt} = -\gamma(t)\nabla \mathrm{E}(\xi) \tag{17}$$

where $\nabla \mathrm{E}(\xi)$ denotes the gradient and $\gamma=[\gamma_{ij}]$ is an $n_p \times n_p$ matrix, n_p is the number of parameters in A and B matrices which are equal to 8 in this case. Therefore, (17) can be written as:

$$\frac{d\xi_i}{dt} = -\gamma \frac{\partial \mathrm{E}}{\partial \xi_i} \quad 1 \le i \le n_p \tag{18}$$

In (18) γ which has been assumed as a constant can be replaced by or absorbed in the nonlinear function:

$$f_i = \beta \frac{1 - e^{\alpha x i}}{1 + e^{-\alpha x i}} \tag{19}$$

so that,

$$\xi_i = f_i\left(-\int \frac{\partial \mathrm{E}}{\partial \xi_i} dt\right) \tag{20}$$

After substituting (15) and (16) into (20) and rearranging we obtain:

$$A_d = f_i\left(-\int\left(A\int\lambda_{dr}\lambda_{dr}^t + B\int i_{ds}\lambda_{dr}^t - \int\dot{\lambda}_{dr}\lambda_{dr}^t + \int\omega_r\lambda_{dr}\lambda_{dr}^t\right)dt\right) \quad (21)$$

$$A_q = f_i\left(-\int\left(A\int\lambda_{qr}\lambda_{qr}^t + B\int i_{qs}\lambda_{qr}^t - \int\dot{\lambda}_{qr}\lambda_{qr}^t + \int\omega_r\lambda_{qr}\lambda_{qr}^t\right)dt\right) \quad (22)$$

$$B_d = f_i\left(-\int\left(A\int\lambda_{dr}\lambda_{dr}^t + B\int i_{ds}\lambda_{ds}^t - \int\dot{\lambda}_{dr}i_{ds}^t + \int\omega_r\lambda_{dr}i_{ds}^t\right)dt\right) \quad (23)$$

$$B_q = f_i\left(-\int\left(A\int\lambda_{qr}\lambda_{qr}^t + B\int i_{qs}\lambda_{qs}^t - \int\dot{\lambda}_{qr}i_{qs}^t + \int\omega_r\lambda_{qr}i_{qs}^t\right)dt\right) \quad (24)$$

These two equations can be expanded as:

$$a_{11} = f\left(\iint\left(-a_{11}\int\lambda_{dr}^2 - a_{12}\int\lambda_{qr}\lambda_{dr} - b_{11}\int i_{ds}\lambda_{dr} - b_{12}\int i_{qs}\lambda_{dr} + \int\dot{\lambda}_{dr}\lambda_{dr} - \int\omega_r\lambda_{qr}\lambda_{dr}\right)dt\right) \quad (25)$$

$$a_{12} = f\left(\iint\left(-a_{11}\int\dot{\lambda}_{dr}^2\lambda_{qr} - a_{12}\int\lambda_{qr}^2 - b_{11}\int i_{ds}\lambda_{qr} - b_{12}\int i_{qs}\lambda_{qr} + \int\dot{\lambda}_{dr}\lambda_{dr} - \int\omega_r\lambda_{qr}^2\right)dt\right) \quad (26)$$

$$a_{21} = f\left(\iint\left(-a_{21}\int\lambda_{dr}^2 - a_{22}\int\lambda_{qr}\lambda_{dr} - b_{21}\int i_{ds}\lambda_{dr} - b_{22}\int i_{qs}\lambda_{dr} + \int\dot{\lambda}_{qr}\lambda_{dr} - \int\omega_r\lambda_{dr}^2\right)dt\right) \quad (27)$$

$$a_{22} = f\left(\iint\left(-a_{21}\int\lambda_{dr}\lambda_{qr} - a_{22}\int\lambda_{qr}^2 - b_{21}\int i_{ds}\lambda_{qr} - b_{22}\int i_{qs}\lambda_{qr} + \int\dot{\lambda}_{qr}\lambda_{qr} - \int\omega_r\lambda_{dr}^2\right)dt\right) \quad (28)$$

$$b_{11} = f\left(\iint\left(-a_{11}\int\lambda_{dr}^2 - a_{12}\int\lambda_{qr}\lambda_{dr} - b_{11}\int\lambda_{dr}^2 + b_{12}\int\lambda_{dr}\lambda_{qr} + \int\dot{\lambda}_{dr}i_{ds} - \int\omega_r\lambda_{qr}i_{qs}\right)dt\right) \quad (29)$$

$$b_{12} = f\left(\iint\left(-a_{11}\int\lambda_{dr}\lambda_{qr} - a_{12}\int\lambda_{qr}^2 - b_{11}\int\lambda_{dr}\lambda_{qr} - b_{12}\int\lambda_{qr}^2 + \int\dot{\lambda}_{dr}i_{qs} - \int\omega_r\lambda_{qr}i_{ds}\right)dt\right) \quad (30)$$

$$b_{21} = f\left(\iint\left(-a_{21}\int\lambda_{dr}^2 - a_{22}\int\lambda_{dr}\lambda_{qr} - b_{21}\int\lambda_{dr}^2 - b_{22}\int\lambda_{dr}\lambda_{qr} + \int\dot{\lambda}_{qr}i_{ds} - \int\omega_r\lambda_{dr}i_{qs}\right)dt\right) \quad (31)$$

$$b_{22} = f\left(\iint\left(-a_{21}\int\lambda_{dr}\lambda_{qr} - a_{22}\int\lambda_{qr}^2 - b_{21}\int\lambda_{dr}\lambda_{qr} - b_{22}\int\lambda_{qr}^2 + \int\dot{\lambda}_{qr}i_{qs} - \int\omega_r\lambda_{dr}i_{ds}\right)dt\right) \quad (32)$$

Based on (25) to (32), a ANN is constructed with the external input to the ANN being $\dot{\lambda}_{dqr}, \lambda_{dqr}, i_{dqs}$ and ω_r, and the internal inputs being the estimated entries of *A* and *B* matrices themselves. Notice that $a_{12} = a_{21} = b_{12} = b_{21} = 0,$ and this observation is used to reduce the complexity of the designed ANN estimator. Where, at the output of

the ANN estimator we have only four outputs, namely, a_1, a_2, b_1, b_2, instead of eight output quantities of which we have four zeroes.

The stator current vector i_{sd}, i_{sq} and ω_r are assumed to be measurable, and the rotor flux vector λ_{dr} and λ_{qr} obtained from the stator voltage model given as follows.

$$\lambda_{dr} = -Li_{ds} + \int \left(v_{ds} - R_s i_{ds}\right) dt \tag{33}$$

$$\lambda_{qr} = -Li_{qs} + \int \left(v_{qs} - R_s i_{qs}\right) dt \tag{34}$$

A three layer network with a total of 31 hard limit neurons is employed to implement the rotor time constant estimator as shown in Fig. 1. The first hidden layer has 16 neurons (tansig activation function neuron with the w_1 and bias θ_1), 7 neurons in the second hidden layer (square activation function neuron with the weight w_2 and bias θ_2), and the output layer has one neuron (linear active function neuron with the weight w_3 and bias θ_3). The network is trained by a supervised method. After 52036 training epochs, the sum squared error arrives at $3.2.10^{-5}$.

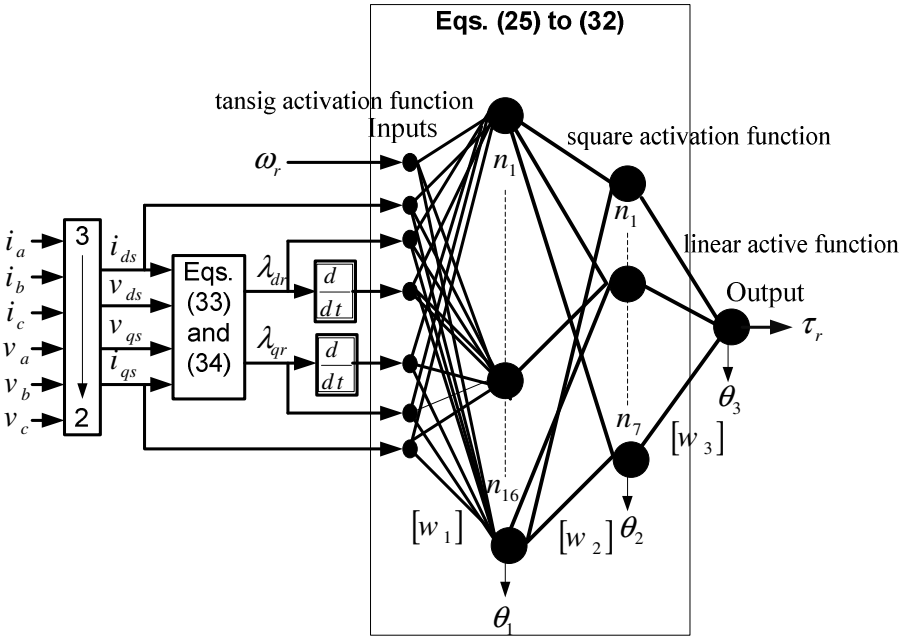
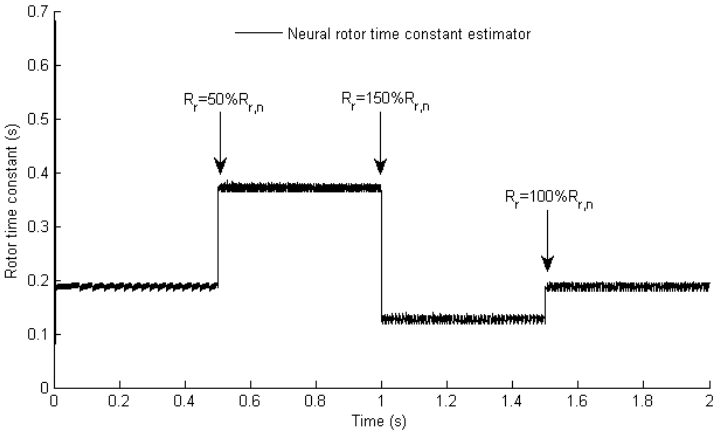


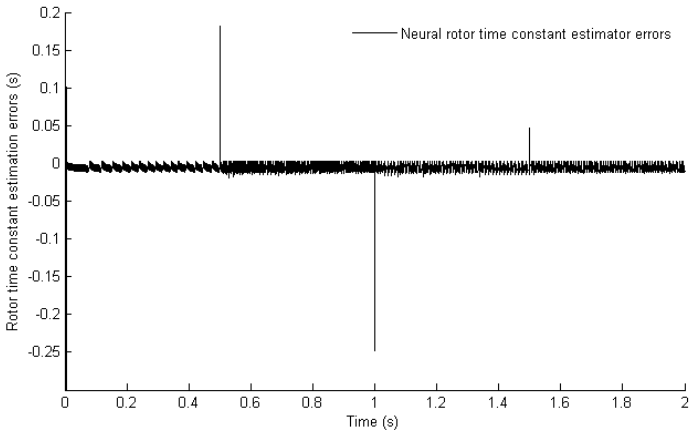
Fig. 1. Structure of neural rotor time constant estimator

Fig. 2 shows the simulation result of neural rotor time constant for the vector controller IM. The rotor time constant estimator was simulated for a reference speed of 300 rad/s. At time t = 0.5s and t = 1s, the rotor resistance (R_r) of the induction motor has a step response which is increased by 50% of R_r and reduced by 150% of R_r

respectively, and at time t = 1.5s, the rotor time constant returns to its normal value. The neural rotor time constant was also used to adjust a rotor flux oriented drive with respect to the rotor resistance variation. The rotor time constant estimated by this ANN is used to correct the set-point slip at vector controller level. However, the rotor time constant change slowly with time change in the real drive. In this simulation, the step variation is to verify the robustness of the proposed estimator. The rotor speed response shows that the drive can follow the low command speed very quickly and rapid rejection of disturbances, with a low dropout speed (Fig. 3). The current responses are sinusoidal and balanced, and its distortion is small (Fig. 4).



(a)



(a')

Fig. 2. (a) Neural rotor time constant estimation; (a') Neural rotor time constant estimation errors

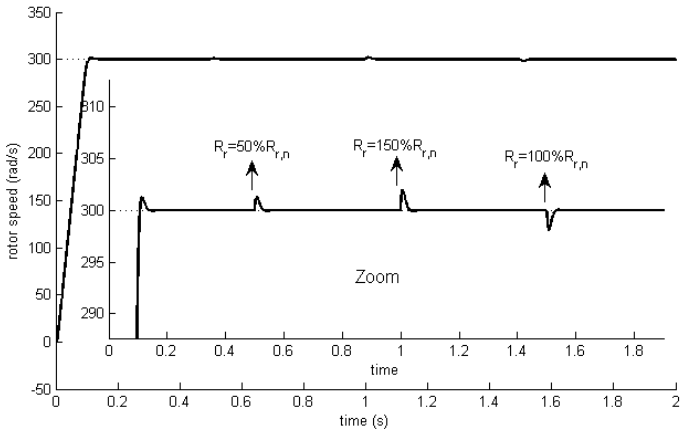


Fig. 3. Rotor speed

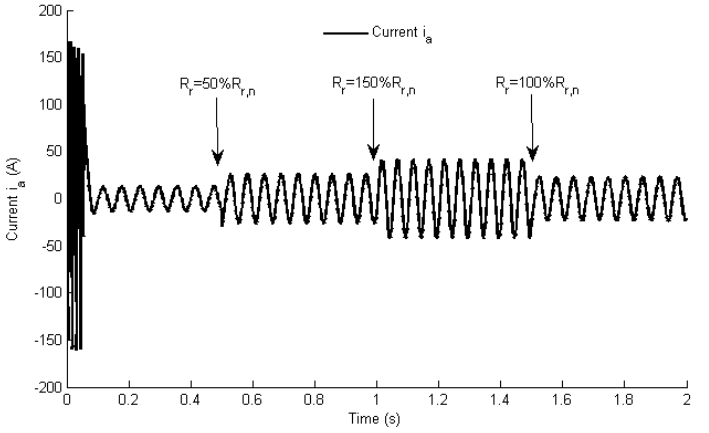


Fig. 4. Current i_a

Induction motor parameters:

P_n = 2.2kW, V_n = 220/380V, f = 60Hz, R_s = 0.84Ω, R_r = 0.3858Ω, L_s = 0.0706H, L_r = 0.0706H, L_m = 0.0672H, J = 0.008kg·m^2, n_p = 2.

5 Conclusion

In this paper we presented the analysis and the discussion of the effect of the rotor time constant variations on the dynamic performance of rotor flux indirect field orientation drives. We have proposed a new method for rotor time constant estimation based on back-propagation neural networks. The computer simulations have shown

the validity and the feasibility of the proposed method that possesses the advantages of neural network implementation: the high speed of processing. In addition this method is more adapted for practical implementation because it uses only stator terminal quantities (voltage, current and frequency) in the estimation of the rotor time constant.

References

1. Leonhard, W.: Control of Electrical Drives, 3rd edn. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, Berlin (2001)
2. Chiasson, J.: Modeling and High-Performance Control of Electric Machines. John Wiley & Sons (2005)
3. Mueller, K.: Efficient Tr estimation in field coordinates for induction motors. In: IEEE International Symposium on Industrial Electronics, vol. 2, pp. 735–741 (1999)
4. Mastorocostas, C., Kioskeridis, I., Margaris, N.: Thermal and slip effects on rotor time constant in vector controlled induction motor drives. IEEE Transactions on Power Electronics 21(2), 495–504 (2006)
5. Wai, R.-J., Liu, D.-C., Lin, F.-J.: Rotor time-constant estimation approaches based on energy function and sliding mode for induction motor drive. Electric Power Systems Research 52(3), 229–239 (1999)
6. Zai, L.C., Lipo, T.A.: Extended Kalman filter approach to rotor time constant measurement in pwm induction motor drives. In: Conference Record - IAS Annual Meeting, pp. 177–183. IEEE Industry Applications Society (1987)
7. Zai, L.-C., DeMarco, C.L., Lipo, T.A.: An extended Kalman filter approach to rotor time constant measurement in PWM induction motor drives. IEEE Transactions on Industry Applications 28(1), 96–104 (1992)
8. Jevremovic, V.R., Vasic, V., Marcetic, D.P., Jeftenic, B.: Speed-sensorless control of induction motor based on reactive power with rotor time constant identification. IET Electric Power Applications 4(6), art. no. IEPAAN000004000006000462000001, 462–473 (2010)
9. Zidani, F., Naït Saïd, M.S., Abdessemed, R., Benbouzid, M.E.H.: A fuzzy method for rotor time constant estimation for high-performance induction motor vector control. Electric Power Components and Systems 31(10), 1007–1019 (2003)
10. Valdenebro, L.R., Bim, E.: Fuzzy optimization for rotor time constant identification of an indirect vector-controlled induction motor drive. In: IEEE International Symposium on Industrial Electronics, vol. 2, pp. 504–509 (1999)
11. Ba-Razzouk, A., Cheriti, A., Olivier, G.: Artificial neural networks rotor time constant adaptation in indirect field oriented control drives. In: PESC Record - IEEE Annual Power Electronics Specialists Conference, vol. 1, pp. 701–707 (1996)
12. Raol, J.R., Madhuranath, H.: Neural network architectures for parameter estimation of dynamical systems. IEE Proceedings: Control Theory and Applications 143(4), 387–394 (1996)

Knowledge Discovery from Heart Disease Dataset Using Optimized Neural Network

R. Chitra[1] and V. Seenivasagam[2]

[1] Department of Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kanyakumari District, India
jesi_chit@yahoo.co.in
[2] Department of Computer Science and Engineering,
National Engineering College, Thoothukudi District, India
yespee1094@yahoo.com

Abstract. Risk Level Prediction at early stage will significantly reduce the risk of Heart Disease. In this paper a novel intelligent technique is proposed to discover the knowledge about the risk of Heart Disease using Optimized Neural Network. A Feed Forward Neural Network optimized using Genetic Algorithm is used for prediction. The network parameters hidden neurons, momentum factor and learning rate are optimized using Genetic Algorithm and the performance is analyzed for standard heart disease dataset and clinical dataset.The classification results prove that the proposed Genetic Optimized Neural Network highly contribute the physician to diagnosis the disease early by discover the knowledge of risk.

Keywords: Genetic Algorithm, Neural Network, Risk Level Prediction, Optimization, Heart Disease.

1 Introduction

To extract hidden patterns from large databases data mining and machine learning techniques are used nowadays. Neural Networks (NN) plays a vital role in the extraction of knowledge from the dataset with proper training. Data Mining and intelligent techniques has been applied for medical diagnosis which needs to entail precise patient data, a philosophical understanding of the medical literature and many years of clinical experience [1] Heart Disease(HD) is a major cause of disability and premature death throughout the world [2]. Many techniques are used to discover the knowledge about the level of risk .Association rule mining based on sequence number and clustering the transactional data base for heart attack prediction was proposed by Jabbar [3]. Risk level of Heart Disease prediction using Decision Tree and Bayesian classification techniques were proposed by K.Srinivas et al[4]. K. Usha Rani et al proposed single and multilayer neural network models to analyze the risk of heart disease and the performance was analysed,it was noticed that the prediction accuracy was high for multilayer feed forward neural network[5]. An intelligent HD prediction system built with the aid of data mining technique like Decision Trees, Nave Bayes

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 423–431, 2013.
© Springer International Publishing Switzerland 2013

and Artificial Neural Network was proposed by Sellappan Palaniappan et al [6]. Asha Gowda Karegowda, et al. [7] diagnoses the diabetes mellitus by using hybrid model of genetic algorithm and back propagation network. In this paper, an intelligent knowledge discovery is proposed using an Optimized Neural Network. The knowledge about the level of risk is identified and the dataset with high risk is classified as abnormal and others are classified as normal in the rest of the section. Backpropagation Algorithm is used for training the feed forward NN. In order to improve the performance of the Feed Forward Neural Network, its parameters are tuned using Genetic Algorithm. The rest of the section is organized as follows. The proposed Artificial Neural Network structure, Genetic Algorithm, Network training and the dataset used for HD prediction is explained in Section 2. In Section 3 HD classification using GA-ANN is discussed. The performance of the developed model is compared with ANN is discussed in Section 4

2 Materials and Methods

In this system the risk of heart disease is more accurately predicted with reduced number of attributes. The ANN parameters are optimized using Genetic Algorithm.

2.1 Heart Disease Dataset

The Datasets used for risk level prediction of HD are the dataset obtained from UCI repository (Cleveland) and real time Clinical Dataset. The attribute and its assigned values for HD prediction are given in table 1.In this work 13 major attributes are used to discover the knowledge about the level of risk. The factors that increase the risk of HD were demonstrated in Framingham in the mid-20th century. HD risk factors never occur in isolation but they are correlated to each other.

2.2 Multilayer Feed Forward Neural Network

Neural Networks are inspired in the biological neural nets and are used for complex and difficult tasks and are capable of generalization and hence the classification is natural for them. The Neural Network is trained from the historical data with the hope that it will discover hidden dependencies and that it will be able to use them for predicting. Feed Forward Neural Networks trained by Backpropagation Algorithm have become a standard technique for classification and prediction tasks[8].

A Multilayer Feed Forward Neural Network is a three layer network with input unit $x_i = \{x_1, x_2, \ldots, x_n\}$, hidden layer $h_j = \{h_1, h_2, \ldots, h_n\}$ and output layer $y_k = \{y_1, y_2, \ldots, y_n\}$. The number of layers in a Neural Network is the number

Table 1. Input Heart Disease Dataset

Attribute	Attribute Values
Age in years	As recorded (lies between 25-75 years)
Gender	value 1: Male; value 0 : Female
Chest Pain Type	value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic
Blood Pressure	As recorded (between 90-192 mm Hg)
Cholesterol	As recorded (between 160-410 mg/dl)
Fasting Blood Sugar	value 1: $\geq$ 120 mg/dl; value 0: $<$ 120 mg/dl
Resting ECG Result	value 0:normal; value 1: ST-T wave abnormality; value 2: definite LV hypertrophy
Maximum Heart Rate	As recorded (between 71-202)
Exercise induced angina	value 1: yes; value 0: no
ST depression induced by exercise	value 1: yes value 0: no
Slope of the peak exercise ST segment	value 1: unsloping; value 2: flat; value 3:downsloping
Number of major vessels (0-3) colored by flourosopy	value 0 - 3
Thallium Test	value 3: normal; value 6: fixed defect value 7:reversible defect

of layers of perceptrons.It consists of a layer of input units one or more layers of hidden units and one output layer of units. The number of input and output units depends upon the application and requires experimentation to determine the best number of hidden units.Each connection between nodes has a weight associated with it. The input layer depends on the number of attributes taken from the patients history (13 attributes). The ANN with13 input attributes 6 hidden neurons and one output neuron is shown in figure 1.

Each hidden node calculates the weighted sum of its inputs and applies a thresholding function to determine the output of the hidden node. The thresholding function applied at the hidden node is a sigmoid activation function.

$$f(x) = \frac{1}{1 + e^{-x}}$$

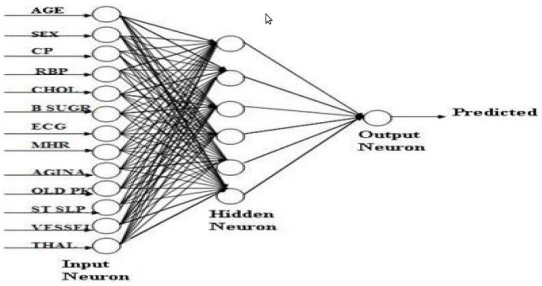


Fig. 1. Proposed ANN Architecture

Training the Neural Network to produce the correct outputs for the given inputs is an iterative process. The network is first initialized by setting up all its weights to be small random numbers between [0, +1]. The 13 inputs are applied and the output is calculated. The actual output (t) obtained is compared with the target output (y) and the error is calculated by finding the difference between the target and the actual output. This error is then used mathematically to change the weights in such a way that the error will get minimized. The process is repeated again and again until the error is minimal. In the proposed work Mean Square Error(MSE) function defined in eqn(1) used for training.

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - t_i)^2 \tag{1}$$

The weights are updated using eqn (2)

$$\triangle w_{ij}(t+1) = -\eta\frac{\partial E}{\partial w_{ij}} + \alpha \triangle w_{ij}(t) \tag{2}$$

In equation (2) α is the momentum factor and η is the learning rate.

The momentum factor α allows for momentum in weight adjustment. Momentum basically allows a change to the weights to persist for a number of adjustment cycles. The magnitude of the persistence is controlled by the momentum factor. Increased momentum factor increasing greater persistence of previous adjustments in modifying the current adjustment. Momentum factor in the BP algorithm can be helpful in speeding the convergence and avoiding local minima.

Learning Rate η is a training parameter that controls the size of weight and bias changes during learning.The selection of a learning rate is of critical importance in finding the true global minimum of the error distance. Increasing number of hidden neurons increasing processing power and system flexibility. But the complexity and cost of the system also increases depends on the number of hidden neurons.

Backpropagation Algorithm

Initialise the weights to small random values between [0, +1]
Initialise the $epochmax = 1000 MSE_{min} = 0.001$
Until satisfied
Do
For each training example (x_i, y_i)
Input the training example and compute the network output
Calculate the MSE
if $MSE \leq MSE_{min}$
Stop training the network
else if $epoch \geq epochmax$
Stop training
else
{Update weights; epoch=epoch+1 }
end
end

Table 2. Optimizing Parameters selected in ANN

Attribute	Attribute Values
Parameter	Range
Number of hidden neurons in hidden layer	6-25
Momentum Factor (α)	0.9-0.931
Learning Rate (η)	0.01-0.114

2.3 Genetic Algorithm

In GA, chromosomes share information with each other. So the whole population moves like a one group towards an optimal area. At each step Genetic Algorithm select the individual randomly and also uses the current population to create the children that can make up the next generation [9]. The evolution starts from a population of randomly generated individuals and in each generation, the fitness of every individual in the population is evaluated, the more fit individuals selected from the current population. The fitness function of a problem is a main source providing the mechanism for evaluating the status , fitness function takes the chromosome as input and produces a number or list of numbers as a measure to the chromosome's performance. GA can be used to improve the performance of the Feed Forward ANN, GA can be optimally designing the ANN parameter which includes the ANN architecture, training algorithm, number of iteration, weights, activation function [10, 11]. In this paper GA has been used to search for the optimal neurons in the hidden layer, training parameter momentum factor and learning rate of ANN for predicting the risk level of HD. In the optimization process the fitness function used is the MSE obtained and Genetic Optimization is introduced in Backpropagation learning .

3 Risk Level Prediction Using GA-ANN

The knowledge about the risk of heart disease is predicted using GA optimized ANN. Initial ANN used for prediction is shown in figure 1 and it is 13-6-1 architecture.Genetic optimization is chosen and applied in Feed Forward Neural Network to enhance the learning process in terms of convergence rate and classification accuracy. In the proposed work the fitness function is the MSE obtained in ANN training given in equation 1. The decision variables used are number of hidden layers, momentum factor and learning rate and all decision variables are bounded within the range specified in table 2.The steps involved in GA optimization is discussed

Step 1: Initialization
All the parameters used in GA are initialized.
Population size =25; Selection operator=Rank based; Crossover probability=0.5
Mutation probability=0.8; Maximum Generation =200.

Step 2: Evaluation of fitness function.
For each generation the MSE is calculated from the output of ANN.

Step 3: Sorting and Rank based selection:
The MSE values of 25 populations are obtained and then obtained value is sorted in ascending order. The first two values are selected Cross over and mutation is performed in the selected offspring with the given probability.

Step 4. Termination
If generation exceeds the maximum generation or MSE lies within the acceptable range terminate the optimization. Otherwise repeat step 2 and 3.

The optimized neural network parameters obtained are momentum factor is 0.922, learning rate is 0.0101 and number of hidden neurons is 20. The ANN is trained in offline hence the time needed to obtain the optimized ANN is ignored during risk level prediction. The Cleveland dataset and the clinical dataset is classified using developed Optimized NN.

4 Result Analysis and Discussion

The performance of the knowledge discovered from Cleveland and Clinical dataset is explained in this section. The performance of the GA-ANN is compared with the ANN to evaluate the sensitivity, specificity and accuracy. The dataset with high risk is considered as normal and other as abnormal for this performance discussion. The Cleveland Dataset has 150 normal and 120 abnormal cases. The clinical data has 300 normal and 200 abnormal cases. Sensitivity evaluate the diagnostic test correctly at detect the positive disease. Specificity measures the diagnostic test correctly at detecting the without disease. Accuracy measured correctly figured out the diagnostic test by eliminating the given condition.

The confusion matrix for Cleveland Dataset and Clinical Dataset are shown in the Table 3.

Table 3. Confusion Matrix

	Cleveland dataset	Clinical Dataset
TP	105	158
FP	15	42
TN	134	269
FN	16	31

Table 4. Error Modeling for Dataset

	Cleveland dataset	Clinical Dataset
FPR (Type I Error)	0.10	0.165
FNR (Type II Error)	0.133	0.136

The type I and type II error obtained in both dataset is listed in Table 4 and from the table it is noticed the error is also in a significantly acceptable range. Performance of the risk level prediction system with ANN and GA-ANN for standard Cleveland Dataset is shown in figure 2. Accuracy of the system is 85.9%, sensitivity 83.6%, specificity 87.8% for the ANN based heart disease prediction and Accuracy 88.5%, sensitivity 86.7%, specificity 89.9% for the GA-ANN based heart disease prediction. Figure 3 shows the performance of the Clinical Dataset using ANN and GA-ANN heart disease prediction system. Accuracy 81.2%, sensitivity 78.4%, specificity 82.8% for ANN algorithm and Accuracy 85.3%, sensitivity 83.5%, specificity 86.4% for GA-ANN algorithm.

The convergence plot for ANN and GA-ANN is shown in figure 4. The mean square is reduced rapidly in GA-ANN compared to ANN. The proposed algorithm yields best convergence than that of the ordinary neural network. The better fitness value obtained in optimization is 0.034.

GA-ANN algorithm is converged in the 101 iteration. But the ANN is converged in the iteration 200 only. ROC curve is a graphical plot created by plotting false positive rate versus true negative rate and it is used to analyze the performance of the classifier. The ROC curve for ANN and GA optimized ANN is shown in figure 5. The area under the curve is maximum for GA-ANN compared to ANN. Hence this model can be used as a diagnostic tool for cardiologist with high precision accuracy compared to ANN. The accuracy of the ANN classifier for HD prediction is 0.856.But in the case of GA-ANN classifier is 0.885.It is proved that the efficiency of the HD prediction system is improved 3% by optimizing the network parameters.

Similarly the sensitivity and specificity of the system is also high for the proposed system compared to existing model. Although ANN produces a good classification accuracy for prediction of HD, this classifier select the initial parameter setting manually which will take a considerable time to classify the patterns. In summary, its found that the proposed GA-ANN based prediction system offers substantial improvement in prediction of HD. This implies that

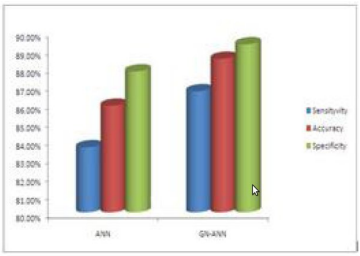


Fig. 2. Performance of Cleveland Dataset

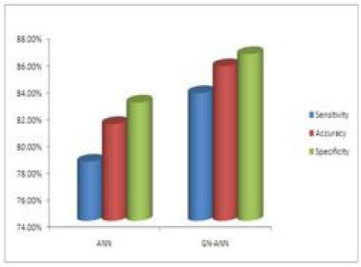


Fig. 3. Performance of Clinical Dataset

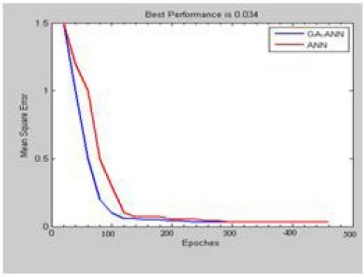


Fig. 4. Convergence Plot

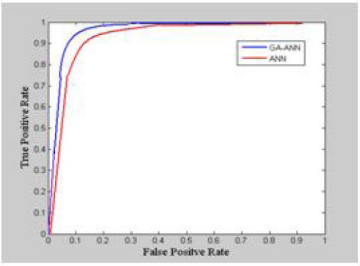


Fig. 5. ROC curve

the optimized ANN is one of the desirable classifier, which can be used as an aid for the physicians to predict the heart diseases.

5 Conclusion

In this paper, we have proposed an adaptive intelligent mechanism for heart disease prediction using Cleveland Dataset and the profiles collected from the patients.GA optimized ANN adopted global and local optimization of the network parameters within the specified range. The developed algorithm is computationally efficient for heart disease prediction .Genetic algorithm can thus evolve an optimum number of hidden units within an architecture space. The results of proposed system, preformed over data sets obtain from 270 Cleveland Database and 500 data collected from patients; shows that it has achieved better accuracy than ANN. There are many interesting aspects for future work. Ant colony optimization can be used to select the input features. Further this system can be enhanced using Swarm intelligence techniques for optimizing the ANN weight.

References

1. Anooj, P.K.: Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University Computer and Information Sciences 24(1), 27–40 (2012)
2. Global status report on no communicable diseases, World Health Organization (2010)
3. Soni, J., Ansari, U., Sharma, D.: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications (0975 8887) 17(8) (2011)
4. Srinivas, K., Raghavendra Rao, G., Govardhan, A.: Survey on prediction of heart morbidity using data mining techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP) 1(3), 14–34 (2011)
5. Usha Rani, K.: Analysis of heart diseases dataset using neural network approach. International Journal of Data Mining & Knowledge Management Process (IJDKP) 1(5), 1–8 (2011)
6. Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques, pp. 108–115. IEEE (2008)
7. Karegowda, A.G., Manjunath, A.S., Jayaram, M.A.: Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes. International Journal on Soft Computing 2(2), 15–23 (2011)
8. Wilamowski, B.W., Chen, Y., Malinowski, A.: Efficient Algorithm for Training Neural Networks with one Hidden Layer. In: Proceedings on the International Conference on Neural Networks, San Diego, CA (1997)
9. Man, K.F., Tang, K.S., Kwong, S.: Genetic Algorithms: Concepts and Applications. IEEE Transactions on Industrial Electronics 43(5) (October 1996)
10. Tian, X., Song, T., Liu, Y.: Improving the structure and the parameter of the BP nerve network with Genetic Algorithm. Journal of Dalian University of Technology 2(6), 69–71 (2004)
11. Weihong, Z., Shunqing, X.: Optimization of BP Neural Network Classifier Using Genetic Algorithm. In: Du, Z. (ed.) Intelligence Computation and Evolutionary Computation. AISC, vol. 180, pp. 599–605. Springer, Heidelberg (2013)

Causality Inference Techniques for *In-Silico* Gene Regulatory Network

Swarup Roy[1], Dipankar Das[1], Dhrubajyoti Choudhury[1], Gunenja G. Gohain[1], Ramesh Sharma[1], and Dhruba K. Bhattacharyya[2]

[1] Dept of IT, North Eastern Hill University, Shillong 793022, Meghalaya, India
swarup@nehu.ac.in, {dipankar.rinku.das,dchoudhury325,gunenja}@gmail.com, r.sharma2905@yahoo.in

[2] Dept of CSE, Tezpur University, Napaam 784028, Assam, India
dkb@tezu.ernet.in

Abstract. Causality detection in gene regulatory networks (GRN) is a challenging problem due to the limit of available data and lack of efficiency in the existing techniques. A number of techniques proposed so far to reconstruct GRN. However, majority of them ignore drawing causality among genes which indicates regulatory relationship. In this paper, we study few techniques available for inferring causality. We select four state-of-the-art causality detection techniques namely, Bayesian network, Granger causality, Mutual information(MI) and Transfer entropy based approach for our study. Performance of the techniques are evaluated using DREAM challenge data based on associated *in-silico* regulatory networks. Experimental results reveal the superiority of MI based approach in terms of prediction accuracy in comparison to other techniques.

Keywords: causality, gene regulatory networks, microarray, prediction, Bayesian, Granger, mutual information, transfer entropy.

1 Introduction

Biological network presents an integrated way to look into the dynamic behaviour of the cellular system through the interactions of components. Biological networks may be categorised [1] as *metabolic pathways*, *signal transduction pathways*, *gene regulatory networks*, *protein-protein interaction(PPI)* [2] networks. Advent of micro-array technology has enabled system biologist to study the dynamic behaviour of genes with respect to different conditions [36]. Due to availability of large collection of microarray data and next generation sequencing technologies, it is now possible to reconstruct or reverse engineer the cellular system *in-silico*.

Gene Regulatory Networks (GRN) is a collection of genes in a cell which interact with each other and with other substances in the cell such as proteins or metabolites, thereby governing the rates at which genes in the network are transcribed into mRNA. Mathematically, GRN can be represented as directed graph, where node represents gene or gene products and edge represents biochemical processes like reaction, transformation, interaction, activation, inhibition.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 432–443, 2013.
© Springer International Publishing Switzerland 2013

In GRN, causal information is one of the important component in inferring regulatory relationship between the genes or gene products. Causation refers to the relation that exists between the cause and its effect, where the effect is an outcome of the cause. It makes a huge contribution in GRN and represented as directed graph. The directed edges in GRNs correspond to causal influences between gene-activities (nodes). These could include regulation of transcription by transcription factors, but also less intuitive causal effects between genes involving signal-transduction or metabolism.

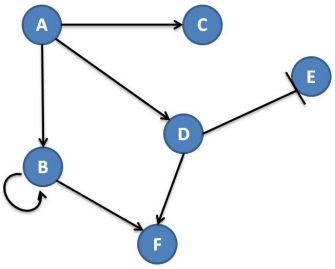


Fig. 1. Possible causal relationship between five nodes in a GRN

A causal effect may be direct or indirect [3]. A gene may influence activity of other gene or gene product directly. On the other hand, a gene may influence activity of other gene or itself by coding a transcription factor (TF) that in turn regulates another gene or itself. A possible causal relationship in GRN is shown in Fig. 1. Apparently, four different types of causal relationship may be possible in a living cell. Based on the above figure we can derive following relationship.

1. A gene can enhance the activity of more than one gene (relationship between A, B, C and D).
2. A gene's activity may be influenced by more than one gene (relationship between B, D and F)
3. Gene can also influence the activity of itself (node B).
4. A gene may inhibit activity of other gene (D inhibit E). Inhibition or negative regulation may also follow above three relationship i.e. many-to one, one-to-many and self.

A number of techniques have been proposed for network construction [4–8, 37]. Many approaches use statistical [38], machine learning or soft-computing techniques [9] as discovery tools. Broadly the techniques can be classified into two categories [10, 11], (i) supervised and (ii) unsupervised approach. Supervised approach requires prior knowledge about the regulatory interactions to infer novel interactions and unsupervised approach infer networks exclusively based on data (e.g. differential gene expression). Most exiting techniques infer regulatory networks without causality information. They depict GRN as an undirected graph, where edge represents some kind of association among the genes

(e.g. correlation or mutual information). Sometime they referred as co-expression network [12, 13]. A few computational techniques are available for reconstruction of GRN that represent GRN as directed graph with causal information. In this work, we review some of the state-of-the-art GRN construction techniques that compute casual relationship between the genes. Prediction accuracy are evaluated using *in-silico* regulatory networks along with associated gene expression data.

2 Causality Detection Techniques

Causality is considered to be fundamentals to natural science like physics, biology and also a topic studied from the perspectives of economics, philosophy and statistics. Inspired from different causality finding techniques applied in statistics, economics or physics, a number of computational techniques proposed so far for reconstruction of GRN with causality. In this work we consider only four benchmark techniques for analysis and comparison. Below we present a brief discussion on all the techniques.

2.1 Bayesian Network

A pioneering work by Friedman et al. [14, 15] introduced Bayesian networks as a probabilistic tool for the identification of regulatory genes using high throughput experimental data [16, 17]. A Bayesian network [18] represents the joint probability distribution of a set of random variables and captures the dependencies and conditional independencies between variables in a graphical manner. A Bayesian network is a representation of a joint probability distribution. This representation consists of two components. The first component, G, is a *directed acyclic graph* (DAG) whose vertices correspond to the random variables $X_1, \cdots, X_n$. The second component, Θ, describes a conditional distribution for each variable, given its parents in G. Together, these two components specify a unique distribution on $X_1, \cdots, X_n$. The graph G represents conditional independence assumptions that allow the joint distribution to be decomposed, economizing on the number of parameters. The graph G encodes the Markov assumption i.e. each variable X_i is independent of its non-descendants, given its parents in G. By applying the chain rule of probabilities and properties of conditional independences, any joint distribution that satisfies Markov property can be decomposed into the product form

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | parent(X_i)). \tag{1}$$

Learning Bayesian Networks: The problem of learning a Bayesian network can be stated as follows. Given a training set $D = \{X^1, \cdots, X^n\}$ of independent instances of X, find a network $B = (G, \Theta)$ that best matches D. The common

approach to this problem is to introduce a statistically motivated scoring function that evaluates each network with respect to the training data and to search for the optimal network according to this score. Score-based learning algorithms are general purpose heuristic optimization algorithms. It searches all the possible DAGs and uses a scoring function to evaluate each stage of the search process. One of the popular score and search based method is the Hill Climbing algorithm [19]. Initially it creates a solution and that solution is the best solution until the heuristic finds a better one. The steps involved in the Hill Climbing techniques are as follows:

1. Initially a solution to the problem is created.
2. Next a small part of the solution is changed to create a slightly different solution. The changes that can be applied to the graph are arc addition, deletion and removal operation. After performing these operations we get a new solution. Now if the new solution is better than the original then the new solution is taken as the current working solution and it is used to create new solutions.
3. If it is not better it is rejected and a different change is made to the original solution.
4. At some point none of the available changes will produce an improvement. When none of the small changes produce an improvement the solution is known as a local optimum. The search process is normally stopped when a local optimum is reached. To mitigate this problem the algorithm is restarted.

Discussion: It is effective in detecting non-linear relationship and easy to recognize the dependence and independence between nodes. Bayesian network is acyclic in nature i.e. there can be no cycle or loop in it. However, in real gene regulatory network loop can not be ignored. Moreover, construction of the network is a NP-hard (nondeterministic polynomial-time hard) problem. The search space increases super-exponentially if the number of variables increases in the network.

2.2 Granger Causality

Clive J Granger introduced the Granger causality tests [20], to analyze the effect of one time series on another one. He thought out of the box and said that 'regressions' does not only show 'correlations' but if certain tests are performed on them they may reveal information about causality. It was then widely used in economics but now a days it has found its application in neuroscience, bioinformatics and some other domains. Granger causality is applied successfully to identify gene-gene interactions from mRNA experiment data to elucidate biological process in disease development. It also used to discover GRN from DNA microarray time-series data [21–24].

Granger causality is a statistical concept of causality that is based on prediction. According to Granger causality, if a signal A "Granger-causes" (or "G-cause") a signal B, then past values of A should contain information that helps

predict future value of B above and beyond the information contained in past values of B alone. This means that A can help in reducing the errors which were inevitable if the calculation was done using the values of B only.

G-causality is normally tested in the context of linear regression models. For illustration, consider a bivariate linear autoregressive model of two variables X_1 and X_2 :

$$X_1(t) = \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t)$$
$$X_2(t) = \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t) \quad (2)$$

where p is the maximum number of lagged observations included in the model (the model order), the matrix A contains the coefficients of the model (i.e., the contributions of each lagged observation to the predicted values of $X_1(t)$ and $X_2(t)$, and E_1 and E_2 are residuals (prediction errors) for each time series. If the variance of E_1 (or E_2) is reduced by the inclusion of the X_2 (or X_1) terms in the first (or second) equation, then it is said that X_2 (or X_1) Granger-(G)-causes X_1 (or X_2). In other words, X_2 G-causes X_1 if the coefficients in A_{12} are jointly significantly different from zero. This can be tested by performing an F-test of the null hypothesis that $A_{12} = 0$, given assumptions of covariance stationarity on X_1 and X_2. The magnitude of a G-causality interaction can be estimated by the logarithm of the corresponding F-statistic.

Discussion: Granger causality is not necessarily true causality. If both X and Y are driven by a common third process with different lags, one might still accept the alternative hypothesis of Granger causality. Yet, manipulation of one of the variables would not change the other. Indeed, the Granger test is designed to handle pairs of variables, and may produce misleading results when the true relationship involves three or more variables.

2.3 Mutual Information Based Technique

Mutual Information is an effective information theoretic measure applied in gene expression for finding biologically significant relationship among genes [25]. However, like correlation, mutual information (MI) along can't able to give causality between two variables. Catharina Olsen et al. used conditional mutual information with MI to successfully infer causal relation in GRN [26]. Mutual information between two events is defined as the information that one event contains about the occurrence of the other event in a particular environment. Mutual information of two variables reacts as the mutual reduction in uncertainty of one by knowing the other one. On the other hand, conditional mutual information is the expected value of the mutual information of two random variables given the value of a third.

The mutual information I(X; Y) between two random variables X and Y is then defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3}$$

and conditional MI between random variables X and Y given Z is defined as:

$$I(X,Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{4}$$

where, $H(X), H(Y), H(X,Y)$ are the measure of uncertainty in the values of the genes X, Y and (X,Y) and conditional entropy $H(X|Y)$ quantifies the amount of information needed to describe the outcome of a random variable X given that the value of another random variable Y is known.

Based on MI and conditional MI, they calculated a interaction score between the variables X, Y and Z as:

$$C(X,Y,Z) = I(X,Y)I(X,Y|Z) \tag{5}$$

where, I(X, Y) is the MI between X, Y and $I(X,Y|Z)$ is conditional MI between X, Y assuming Z has already occurred.

Catharina Olsen et al. used the concept of *V-structure* in their work. If in a network three nodes *X, Y, Z* is connected as $X \rightarrow Y \leftarrow Z$, then B is called a *collider* and the structure is called a *v-structure*. The negative value of the interaction information $C(X,Y,Z)$, implies Z is a collider.

Following are the steps involved in above approach:

1. Inference of the undirected network using a method able to cope with the usually high number of variables (e.g. ARACNE [8], CLR [27] or MRNET [28]).
2. Estimate the interaction information for all possible v-structures.
3. Use the interaction information criterion to orient the v-structures.

By using the undirected graph obtained in the first step, the interaction information for all possible v-structures is measured. Colliders are detected based on negative value of interaction information. In the third step, all the v-structures are orientated. The orientation can be done by arranging the valid values calculated in the previous step in random order, in decreasing order, or in decreasing order of average interaction information values which are undetectable by the MI based technique.

Discussion: The graph obtained in this approach is a mixed graph, containing both the directed and the undirected edges, which is an important component in real GRN. The MI based approach works well only with v-structure in a graph. As shown in Fig. 1, other then v shaped structure a number of other structures are also available in a GRN.

2.4 Transfer Entropy Based Technique

Schreiber [29] developed the concept of Transfer Entropy between two processes to capture the nonlinearities that are common in real-world system. The Transfer

Entropy is developed from the concept of Kullback entropy [30]. It measures the uncertainty reduction in inferring the future state of a process by learning the current and past states of other processes. The transfer entropy is the amount of information flow from one process to the other. Originally applied in Physics, transfer entropy later applied in inferring genetic networks [31, 32].

Let X_i and Y_i be two time series with discrete states x_i, y_i at time i. Assume that the series can be approximated by a stationary Markov process of order k. Then the dynamical structure of that process is reflected by the transition probability $p(x_{i+1}|X_i^{(k)})$. The true transition probability is usually not known so that one has to assume a prior transition probability $q(x_{i+1}|X_i^{(k)})$.

Suppose that the future state x_i of X_i depends on k past states of X_i but not on the l past states of Y_i then the generalized Markov property holds:

$$p(x_{i+1}|X_i^{(k)}, y_i^{(l)}) = p(x_{i+1}|X_i^{(k)}) \tag{6}$$

If there is a dependence of X on Y, it can be quantified by Kullback entropy with $p(x_{i+1}|X_i^{(k)}, y_i^{(l)})$ as underlying transition probability and $p(x_{i+1}|X_i^{(k)})$ as a prior transition probability. The amount of information transferred from Y to X is defined using Transfer Entropy as follows:

$$\mathrm{T}_{Y\to X} = \sum_{x_{i+1}, X_i^{(k)}, y_i^{(l)}} p(x_{i+1}, X_i^{(k)}, y_i^{(l)}) \log \frac{p(x_{i+1}|X_i^{(k)}, y_i^{(l)})}{p(x_{i+1}|X_i^{(k)})} \tag{7}$$

By assessing the transfer entropy between all pair of genes one can infer a causal network of genes and then apply a heuristic rule to differentiate indirect and direct causal relations. Tung et al. [32] proposed a three steps method for reconstruction of gene regulatory network from microarray time series data. The steps are as follows:

1. Quantify causality relations between all pair of genes by measuring the transfer entropy between their time series data.
2. Estimate the significant levels of all causality relations and select ones whose significant level is greater than a predefined threshold value. The selected relations are used to construct a directed graph.
3. Refine the graph by identifying and removing edges which are consider as indirect causal relations.

Discussion: The transfer entropy cannot only identify the linear causality but also nonlinear causality. It also requires a low computational effort and applicable to a large-scale analysis because it is a model-free network reconstruction method which is based on pair wise statistical measure of causal relations. Along with the advantages discussed there are some disadvantages of transfer entropy. The causal relationships inferred by transfer entropy are often misleading when the underlying system contains indirect connections, dominance of neighboring dynamics, or anticipatory couplings.

Table 1. Properties of different causality inference techniques

	Bayesian Network	**Granger Causality**	**Mutual Information**	**Transfer Entropy**
Nodes	Random variables	Random variables	Random variables	Random variables
Edges	Joint probability distribution	Linear regression	Conditional MI	Transition probability
Input Parameters	Multivariate	Bivariate	Multivariate	Bivariate
Causality	Non linear	Linear	Linear	Non linear

The overall properties of the four techniques are summarized in Table 1.

Next, we evaluate all the four causality finding techniques experimentally in terms of accuracy of GRN prediction.

3 Performance Evaluation

In this section we compare the strength of the candidate causality inference techniques for GRN against prediction accuracy. We use *in-silico* gene regulatory networks and associated gene expression data from DREAM (Dialogue for Reverse Engineering Assessments and Methods) network inference challenge, provided by Marbach's Dream Net Weaver [33] platform. Dream3 and Dream4 are the two challenges that are available. Dream3 involves fifteen benchmark datasets, five each of various sizes (10, 50 and 100). The structures of the benchmark networks are obtained by extracting modules from real biological networks. At each size, two of the networks are extracted from the regulatory network of E. coli and Yeast. Dream4 is very similar to Dream3 containing a total of 10 networks, five of each size, 10 and 100. The *in silico* datasets generated based on [33] platform for our experiments are characterized in Table 2.

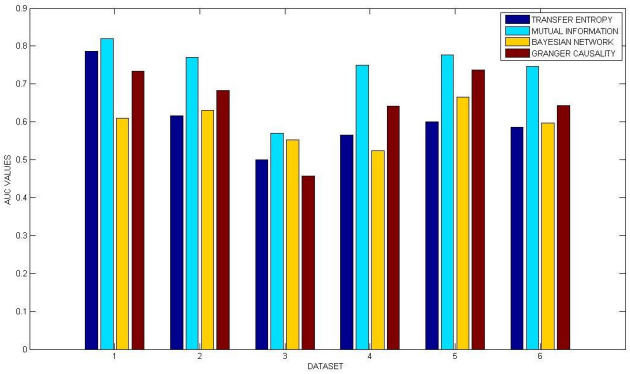
Table 2. *In silico* DREAM Challenge datasets

Challenges	Dataset	*In silico* network	Size of the network
Dream3	1	Ecoli1	10
	2	Ecoli2	10
	3	Yeast1	50
	5	Yeast2	50
Dream4	4	insilico1	10
	6	insilico2	10

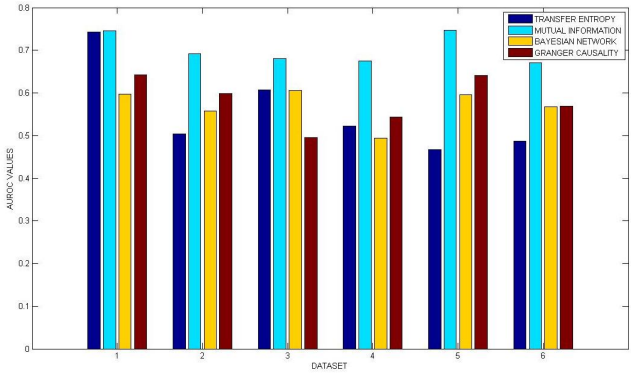
Prediction effectiveness is compared against the actual networks generated from *in slico* DREAM Challenge data, using two different metrics for evaluating accuracy: AUPvR (Area under Precision vs Recall curve) and AUROC (Area under Receiver-Operator Characteristics curve) score. The ROC is also known

as a relative operating characteristic curve, because it is a comparison of two operating characteristics (True Positive Rate and False Positive Rate) as the criterion changes [34]. ROC curves may not be the appropriate measure when a dataset contains large skews in the class distribution, which is commonly the case in transcriptional network inference. As an alternative, precision vs. recall (PvR) curves are considered for measuring prediction accuracy [35]. ROC curves are commonly used to evaluate prediction results. However, PvR curve may be more sensitive when there is a much larger negative set than positive set. Computing the area under the curve (AUC) of a ROC or PvR is a way to reduce ROC or PvR performance to a single value, representing expected performance. The effectiveness of prediction by the four techniques on all the datasets are shown in Fig. 2.

From the figure (Fig. 2) it is evident that MI based technique outperforms rest three candidate techniques in terms of network prediction on two different scores. In case of dataset 1, MI based technique achieved a very high AU(PvR) score of .82 and AUROC of .74. In all cases performance of Bayesian network



(a) AU(PvR) curve of different algorithms



(b) AUROC curve showing prediction performance

Fig. 2. Performance comparison of four causality inference techniques on *in silico* dataset

are not effective. Transfer entropy and Granger causality performs variably for different datasets.

4 Conclusion

In this work we study and analyze four techniques such as Bayesian network, Granger causality, Mutual information and Transfer entropy for inferring causality among the genes to form a Gene Regulatory Network. Based on experimental results it is evident that the Mutual information based techniques yields significantly better results than the other techniques. Majority of the techniques for drawing causal relationship are not biologically motivated. A gene may up- or down-regulate another gene in a network which is missing in the above techniques. Moreover, they consider only few network structures from real GRN. A new causality inference technique which is biologically motivated and draw all the relationship between the genes in a network (Fig. 1) is a current issue in gene regulatory network inference research.

References

1. Tavazoie, S., Hughes, J., Campbell, M., Cho, R., Church, G., et al.: Systematic determination of genetic network architecture. Nature Genetics 22, 281–285 (1999)
2. Panchenko, A., Przytycka, T.: Protein-protein interactions and networks: identification, computer analysis, and prediction, vol. 9. Springer (2008)
3. Brazhnik, P., de la Fuente, A., Mendes, P.: Gene networks: how to put the function in genomics. TRENDS in Biotechnology 20(11), 467–472 (2002)
4. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. J. of Computational Biology 7(3-4), 601–620 (2000)
5. Davidich, M., Bornholdt, S.: Boolean network model predicts cell cycle sequence of fission yeast. PLoS One 3(2), e1672 (2008)
6. Segal, E., et al.: Rich probabilistic models for gene expression. Bioinformatics 17(suppl. 1), S243–S252 (2001)
7. Kuo, W., et al.: Functional relationships between gene pairs in oral squamous cell carcinoma. In: Proc. of AMIA Symposium, pp. 371–375 (2003)
8. Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., Califano, A.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(suppl. 1), S7 (2006)
9. Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(1), 94–107 (2011)
10. De Smet, R., Marchal, K.: Advantages and limitations of current network inference methods. Nature Reviews Microbiology 8(10), 717–729 (2010)
11. Ambroise, J., Robert, A., Macq, B., Gala, J.L., et al.: Transcriptional network inference from functional similarity and expression data: a global supervised approach. Statistical Applications in Genetics and Molecular Biology 11(1), 1–24 (2012)
12. Fuente, A.D.L.: What are gene regulatory networks? In: Handbook of Research on Computational Methodologies in Gene Regulatory Networks, pp. 1–27 (2010)
13. Lee, H., Hsu, A., Sajdak, J., Qin, J., Pavlidis, P.: Coexpression analysis of human genes across many microarray data sets. Genome Research 14(6), 1085–1094 (2004)

14. Friedman, N.: Inferring cellular networks using probabilistic graphical models. Science 303(5659), 799–805 (2004)
15. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology 7(3-4), 601–620 (2000)
16. Dondelinger, F., Husmeier, D., Lèbre, S.: Dynamic Bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series. Euphytica 183(3), 361–377 (2012)
17. Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1), 71–79 (2005)
18. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
19. Gámez, J.A., Mateo, J.L., Puerta, J.M.: Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Mining and Knowledge Discovery 22(1-2), 106–148 (2011)
20. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, 424–438 (1969)
21. Zhang, Z.G., Hung, Y.S., Chan, S.C., Xu, W.C., Hu, Y.: Modeling and identification of gene regulatory networks: a granger causality approach. In: 2010 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 6, pp. 3073–3078. IEEE (2010)
22. Mukhopadhyay, N.D., Chatterjee, S.: Causality and pathway search in microarray time series experiment. Bioinformatics 23(4), 442–449 (2007)
23. Nagarajan, R., Upreti, M.: Granger causality analysis of human cell-cycle gene expression profiles. Statistical Applications in Genetics and Molecular Biology 9(1) (2010)
24. Tam, G.H.F., Chang, C., Hung, Y.S.: Gene regulatory network discovery using pairwise Granger causality. IET Systems Biology 7(5), 195–204 (2013)
25. Priness, I., Maimon, O., Ben-Gal, I.: Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics 8(1), 111 (2007)
26. Olsen, C., Meyer, P.E., Bontempi, G.: Inferring causal relationships using informationtheoretic measures. In: Proceedings of the 5th Benelux Bioinformatics Conference, BBC 2009 (2009)
27. Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J., Gardner, T.: Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biology 5(1), e8 (2007)
28. Meyer, P., Kontos, K., Lafitte, F., Bontempi, G.: Information-theoretic inference of large transcriptional regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology 2007 (2007)
29. Schreiber, T.: Measuring information transfer. Physical Review Letters 85(2), 461 (2000)
30. Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley & Sons (2012)
31. Banerji, C.R., Severini, S., Teschendorff, A.E.: Network transfer entropy and metric space for causality inference. Physical Review E 87(5), 052814 (2013)
32. Tung, T.Q., Ryu, T., Lee, K.H., Lee, D.: Inferring gene regulatory networks from microarray time series data using transfer entropy. In: Twentieth IEEE International Symposium on Computer-Based Medical Systems, CBMS 2007, pp. 383–388. IEEE (2007)

33. Marbach, D., Schaffter, T., Mattiussi, C., Floreano, D.: Generating realistic in silico gene networks for performance assessment of reverse engineering methods. Journal of Computational Biology 16(2), 229–239 (2009)
34. Swets, J.A.: Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. Lawrence Erlbaum Associates, Inc. (1996)
35. Craven, J.: Markov networks for detecting overlapping elements in sequence data. In: Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference, vol. 17, p. 193. MIT Press (2005)
36. Roy, S., Bhattacharyya, D.K., Kalita, J.K.: CoBi: Pattern based co-regulated biclustering of gene expression data. Pattern Recognition Letters 34(14), 1669–1678 (2013)
37. Roy, S., Bhattacharyya, D.K.: Reconstruction of genetic networks in yeast using support based approach. In: Trendz in Information Sciences & Computing (TISC), pp. 116–121. IEEE (2010)
38. Roy, S., Bhattacharyya, D.K.: Mining strongly correlated item pairs in large transaction databases. International Journal of Data Mining, Modelling and Management 5(1), 76–96 (2013)

Multidimensional Longest Increasing Subsequences and Its Variants Discovery Using DNA Operations

Balaraja Lavanya* and Annamalai Murugan

Department of Computer Science, University of Madras,
Chennai, India
lavanmu@gmail.com

Abstract. The Multidimensional Longest Increasing Subsequence (MLIS) and Multidimensional Common Longest Increasing Subsequence (MCLIS) have their importance in many data mining applications. This work finds all increasing subsequences in n sliding window, longest increasing sequences in one and more sequences, decreasing subsequences and common increasing sequences of varied window sizes with one or more dimensions. The proposed work can be applied to finding diverging patterns, constraint MLIS, sequence alignment, find motifs in genetic databases, pattern recognition, mine emerging patterns, and contrast patterns in both, scientific and commercial databases. The algorithms are implemented and tested for accuracy in both real and simulated databases. Finally, the validity of the algorithms are checked and their time complexity are analyzed.

Keywords: LIS, MLIS, MLDS, CMLIS, Pattern recognition, Molecular computing.

1 Introduction

We consider the problem of extracting a Multidimensional Longest Increasing Subsequence (MLIS) from a sequence of integers. The sequence S is assumed to be a permutation of the set 1, 2,$\cdots$, n, with its own dimensions. But having multiple occurrences of integers between 1 and n, in the sequence of length n, does not change the result. Efficiently searching for multidimensional substrings or generally different patterns in large databases is needed today. Genetic codes are stored in DNA molecules. The method of subsequence searching should be insensitive of random insertions, deletions. The nature of identifying patterns varies with applications. The concern is also on the quality of identified patterns. The time taken to discover them plays a vital role in huge researches. These prime issues motivates the proposed work.

A detailed survey of several multiple-string alignment algorithms can be found in [11]. They encountered many notable problems like, the task of optimally

* Corresponding author.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 444–452, 2013.
© Springer International Publishing Switzerland 2013

aligning a set of strings is computationally very expensive [14]. To overcome the difficulty of alignment problem, modified Position Weight Matrices (PWM) [4] can be used to focus on the positions of the patterns in the sequences. Using DNA strands and DNA operations [2], the storage and retrieval of data can be done parallely, this reduces the time and space complexity. In bioinformatics, the two predominant applications of motif discovery are sequence analysis and micro array data analysis. The definition of the search problem, especially the formulation of objective functions, leaves space for substantial improvement in the performance of the motif discovery tool [15].

2 Literature Review

The MLIS problem is closely related to the longest increasing subsequence and longest common subsequence problem, which has a quadratic time dynamic programming solution [8]. Algorithms for finding the LIS date back to Robinson [12] and Schensted [7] with a generalization by Knuth [9]. Fredman [13] showed how to compute an LIS of a length n sequence in optimal O(n log n) time. When the input sequence is a permutation of $\{1, \cdots, n\}$, Hunt and Szymanski [17] designed an O(n log log n) time solution, which was later simplified by Bespamyatnikh and Segal [16]. The support of a pattern is the number of sequences containing the given pattern and its commonality between various other sequences. The longest increasing subsequence problem refers either to identifying the longest increasing subsequence(s) or, alternatively, to determining the length k of the LIS. In either of these forms, this problem has been the subject of intense study by mathematicians and computer scientists alike. This problem has interesting properties both from a purely combinatorial perspective, as well as actual applications in fields such as DNA sequence matching [3].

Simulation of all the DNA operations are done in [2], the proposed work uses *cut* and *pcr* DNA operations. Mining GCS, using DNA operations and modified PWM, given a sequential database is performed in [1]. The concept of multidimensional sequence mining with modified PWM is done in [6]. Discovery of longest increasing subsequences of any window size, with given constraint, from one or more sequences is done in [5]. This paper is an extension of [5] and discovers MLIS.

3 Definitions

Definition 1 *(Dimension): A table T is a set of tuples* $< D, A_1, A_2, A_3, ..., A_N >$*, where D is an attribute whose domain is totally ordered. A sequence S is denoted by an ordered list* $< t_1, t_2, t_3, ..., t_k >$*, where* t_1 *is a tuple, i.e* $D(t_1) \leqslant D(t_2) \leqslant D(t_3) \leqslant D(t_k)$ *for* $1 \leqslant i \leqslant k$*, that is,* $D(t_i)$ *= value of D tuple* t_i*,* $1 \leq i \leq k$*. Every tuple has n analysis attributes along with an ordered value. A set of analysis attribute values can occur at most once in a same value of D, but can occur multiple times in different values of D attribute. The number of values of D in a sequence is called the length of the sequence. If the length of S in k , then we call it a k -sequence [10].*

A multidimensional sequence database is of schema $< D,\ A_1, A_2, A_3, ..., A_N, R_1, ..., R_m >$, where R_i are called relevant dimensions. The schema is partitioned according to relevant attribute values and support computed by number of partitions that contain sequence. As mentioned previously, every partition is similar to table T, a set of tuples $< D, A_1, A_2, A_3, ..., A_N >$ [10]. Given a minimum support threshold min-support, a multidimensional sequence S is called a multidimensional sequential pattern if and only if support $(S) \geq min - support$.

Definition 2 *(Longest Increasing Subsequence): Given a sequence* $S = (s_1, s_2, ..., s_n)$ *and a window size* $w \leqslant n$*, a window of width* w *is a subsequence* $(s_{i+1}, s_{i+2}, ..., s_{i+w})$ *for some* $0 \leqslant i \leqslant n - w$*. We also consider the truncated windows* $(s_1, ..., s_j)$ *for* $j \leqslant w$ *and* $(s_j, ..., s_n)$ *for* $j \geq n - w$ *as windows of size* w*. The general problem that we consider is that of determining a LIS in each of the windows* w_i *and also termed as Longest Increasing Subsequence in Sliding Window (LISSW). If the size of* w *is fixed it is termed as Longest Increasing Subsequence in Fixed Window (LISFW).*

No-of-Elements	{ 1,2,3,4,5,6,7,8,9 }
Input Sequence	6 9 8 2 3 5 1 4 7
All IS	[1 4 7 / 2 3 4 7 / 2 3 5 7 / 3 4 7 / 3 5 7 / 4 7 / 5 7 / 6 9 / 6 8 / 6 7]
LIS	2 3 4 7 / 2 3 5 7

Fig. 1. LIS for each of the given element

4 DNA Based MLIS and Its Variant Patterns Discovery

Algorithms 1 and 2 searches for all increasing sequences with dimensions of different window sizes and different other variant patterns, in one and more input sequences, using modified Position Weight Matrix (PWM). Our approaches makes minimal assumptions about the background sequence model and the mechanism by which elements affect gene expression.

4.1 Finding MLIS in Single Sequence for all Window Sizes (MLISW)

Algorithm MLISW discovers all increasing subsequences, LIS, MLIS and different related patterns, using DNA operations and modified PWM, for different window sizes with single dimension, as shown in Figure 2.

Let $no\_of\_elements$ (noe), be the set of elements, such that $S = (s_1, s_2, s_3, ..., s_m) \in (noe)$, and $window\_size$ be the set of window sizes starting from 2 to $max(S)$. The output of Algorithm 1 is $LISW$ strands for window size 2 to $max(window\_size)$ and $MLISW$ strand.

Algorithm 1. DNA-based-MLIS discovery in single sequence, with one dimension, for all Window sizes (MLISW)

Input: S, $no\_of\_elements$ (noe), $window\_size$, $dimension1[]$
Output: $LISW$ and $MLISW strands$

```
1  begin
2      let n ← max(noe);
3      let m ← max(window_size);
4      let t_1 ... t_n ← pcr(S);
5      PWM_1 ← cut(t_1 , noe[1]) ;
       ⋮
6      PWM_n ← cut(t_n , noe[n]) ;
7      [parallely for each window size LISW_2,LISW_3,...,LISW_m] ;
8      foreach windowsize from 2 to m do
9          [Create threads parallely] ;
10         foreach i from 1 to |S| do
11             if PWM_1[i] > 0 then
12                 test ← PWM_1[i] ;
13                 foreach j from i + 1 to |S| do
14                     if (PWM_1[j] > test) then
15                         LISW_2[k][0] ← i ;
16                         LISW_2[k][1] ← PWM_1[j];
17                     end
18                 end
19             end
20         end
21     end
22     foreach windowsize from 2 to m do
23         [Create threads parallely] ;
24         foreach LISW_2[][] > 0 do
25             let k ← dimension1[LISW_2[][]];
26             if (LISW_2[][] == k) then
27                 MLISW_2[] ← LISW_2[][] ;
28             end
29         end
30     end
31     Extended for higher window sizes and dimensions ;
32 end
```

Steps 1 to 21 of algorithm 1 depicts steps for finding IS, for window size 2, which could be extended for any window sizes each done parallely [5]. The contents of $LISW_m$ are the required LIS of the given S. Steps 22 to 30 performs the check for the specified inequality (equal or ascending order of descending order dimension) parallely and stores the result in $MLISW$ strand. This algorithm can be used for finding Shortest Increasing Subsequences (SIS). The minimum the window size, the shorter is the discovered subsequence length, thus algorithm 1, finds SIS also.

```
MLIS with decreasing order dimension
Input Sequence    6 9 8 2 3 5 1 4 7
Dimension         9 8 7 6 5 4 3 2 1
Window size (W)   { 2, 3, 4, 5}
   W = 2    6 9 / 6 8 / 6 7 / 2 3 / 2 5 / 2 4 / 2 7 /
            3 5 / 3 7 / 3 4 / 5 7 / 1 4 / 1 7 / 4 7
   W = 3    2 3 5 / 2 3 4 / 2 3 7 / 2 5 7 /
            2 4 7 / 3 5 7 / 3 4 7 / 1 4 7
   W = 4    2 3 5 7 / 2 3 4 7
   W = 5    ----------
```

Fig. 2. MLIS for all window sizes with decreasing dimension

Time Complexity The time complexity of Algorithm 1 can be analyzed in 2 phases. The phase 1 is discovery of LIS [6] and discovery of MLIS from LIS constitutes phase 2. Therefore,
TC(MLISW) = O(LIS) + O(MLIS).
If $PWM \notin \emptyset$, then
$TC(LIS)$ is between $(O(n/M) + O((n/L) + n))$ and $O((n-1)(n-1)!)$ at its average case [5]
$TC(MLIS) = O(|LIS|)$
If $|LIS| \in \emptyset$,
$TC(LISW) = \max((O(n/M)+O((n/L)+n))$ and $O((n-1)(n-1)!), O(|LIS|))$ *at its average case*. In the worst case the Algorithm 1, acts as a sequential algorithm. □

4.2 Finding MLIS for Each of the Given Element (MLISGE)

Algorithm MLISGE discovers all increasing sub sequences of, each of the given elements and its variant patterns, in a given sequence using DNA operations, as shown in Figure 1.

Let *no_of_elements* (*noe*), be the set of elements, such that $S = (s_1, s_2, s_3, \ldots, s_m) \in (noe)$. The output of Algorithm 2 is $allLIS$ and $MLIS$ strands for all $noe[1], noe[2], \ldots, noe[n]$.

Steps 2 to 15 of algorithm 2 depicts finding all increasing sub sequences for first element of *noe*, by vertically checking the contents of PWM_1, with all

Algorithm 2. DNA-based-MLIS discovery for each of the Given Element (MLISGE)

Input: S, $no\_of\_elements(noe)$
Output: $allLIS$ *and* $MLIS$ *strand*

```
1  begin
2      let n ← max(noe);
3      let t_1 ... t_n ← pcr(S);
4      let f, s, j, z1 ← 0 ;
5      PWM_1 ← cut(t_1 , noe[1]) ;
           ⋮
6      PWM_n ← cut(t_n , noe[n]) ;
7      [Parallely foreach of the element in noe];
8      foreach iranges from 1 to noe[1]...noe[n]  PWM_i[j] > 0 do
9          foreach  j ranges from i + 1 to noe[1]...noe[n] do
10             if (PWM_i[j] < PWM_z1[j]) then
11                 allLIS_2[f][s] = PWM_z1[j] ;
12                 [Increment f and s] ;
13             end
14         end
15     end
16     [Parallely for each of the element in noe] ;
17     foreach s ranges from 1 to allLIS_2[][] > 0 do
18         [Parallely for each Inequality] ;
19         let k ← dimension1[allLIS_2[][]];
20         foreach  i ranges from s to allLIS_2[][] > 0 do
21             if (dimension1[allLIS_2[][]] == k) then
22                 MLIS_2[] = allLIS_2[][];
23             end
24         end
25     end
26 end
```

other PWM_2 ,..., PWM_n. Steps 17 to 25, checks for the required inequality of dimensions in the discovered LIS and stores the result in $MLIS$ strand. Thus finding all increasing subsequences of each of the element of *noe*,thus algorithm 2, also finds LIS for each of element of *noe*. Similarly, this algorithm can be used for finding Shortest Increasing Subsequences (SIS) for all elements of *noe*.

Time Complexity. Like Algorithm 1, the time complexity of Algorithm 2 can be analyzed in 2 phases. The phase 1 is discovery of LIS [6] and discovery of MLIS from LIS constitutes phase 2. Therefore,
$TC(MLISGE) = O(LIS) + O(MLIS)$

If $PWM \notin \emptyset$, then
$TC(LIS)$ is between $(O(n/M) + O((n/L) + n))$ and $O((n-1)(n-1)!)$ at its average case [5]
$TC(MLIS) = O(|LIS|)$
$TC(MLISGE) = max((O(n/M) + O((n/L) + n))andO((n-1)(n-1)!), O(|LIS|))$ *at its average case.* □

Input Sequence	6 9 8 2 3 5 1 4 7
Dimension	2 1 3 4 5 6 1 2 3
Fixed Window Size = 3	
Sequences are checked as	6 9 8 -- 9 8 2 -- 8 2 3 -- 2 3 5 -- 3 5 1 -- 5 1 4 -- 1 4 7
Increasing Sub sequences	6 9 / 2 3 5 / 3 5 / 1 4 7
Longest Increasing Sub sequences with ascending order dimension	2 3 5 / 1 4 7 4 5 6 / 1 2 3

Fig. 3. Example of MLISSW and MLISFW

4.3 Variants of MLIS

There are many variants of LIS, depending on its application. Algorithm 1 and Algorithm 2 can be used to find some of the variants listed below.

Special Case 1 : Finding MLISSW, MLISFW, CMIS and LCMIS. Multidimensional Longest Increasing Subsequence in Sliding Window (MLISSW), or Multidiemnsional Longest Increasing Subsequence in Fixed Window (MLISFW), can be discovered with algorithms 1, 2 , see Figure 3 . Since algorithm 1 finds increasing sequences for all given window sizes, it can be modified to find MLIS with sliding window also. The step 13 of algorithm 1 can be modified as j ranging from $i+1$ to $windowsize$. Hence find Common Multidimensional Increasing Subsequence (CMIS), and Longest Common Multidimensional Increasing Subsequence (LCMIS).

Special Case 2 : Finding MDS, MCDS, HIS and HCIS. Algorithms 1, 2 can be used for finding Multidimensional Decreasing Subsequence (MDS) and Multidimensional Common Decreasing Subsequence (MCDS) in a single or more number of sequences. In step 14 of Algorithm 1, and step 10 of algorithm 2, the relational operator used to discover increasing subsequences have to be changed to find DS, thereby finding CDS. With a minimum modification, Heaviest Increasing Subsequence (HIS) and Heaviest Common Increasing Subsequence (HCIS) with dimensions can also be found using Algorithm 1 and 2. This needs an additional DNA strand to store each element's respective weight.

5 Performance

Algorithms 1, 2 have been implemented and tested with simulated and real databases. The random DNA sequences of size varying from 100 to 25000, are generated from $http : //old.dnalc.org/bioinformatics/dnalc - nulceotide - analyzer.htm\#randomizer$ and $http : //old.dnalc.org/bioinformatics.org/sms/rand - dna.html$. The real data is collected from $EMBL$ database in $FASTA$ format. The genome sequences of 3021 viruses are collected and tested for the existence of all required patterns. The database is got from $http : //www.ebi.ac.uk/genomes/virus.html$. Tested with randomly generated and real motifs, our work could discover all motifs present, with its positions of existence. All implementations are performed on a dual core computer and 5 GB main memory using Java. The operating system is Windows XP. The resulted data of these experiments are consistent. The limitation of these algorithms is that the maximum number of threads generated, is dependent on the efficiency of the system architecture.

6 Applications

The assumption behind the discovery of patterns is that a pattern that appears often enough, in a set of biological sequences, is expected to play a role in defining, the respective sequences functional behavior and evolutionary relationships. Since the proposed new algorithms use DNA strands for its DNA operations and other processing, the storage and retrieval processes can be implemented easily and parallely, whatever may be the size of the database. The searching for LIS and CLIS and all its variants, has its importance in many industrial, research and scientific applications. Especially in medical and genetic field, the finding of all patterns of motifs with its diverging pattern, can be used to predict, analyse, interpret and conclude the existence or future liability of any disease or abnormality present in the patient data or defaulters in any commercial databases.

7 Conclusion

In this paper, we have designed and performed the implementations to find LIS, MLIS, MCLIS, and different variants of it, in a highly parallel way, and can be extended to many other data mining applications also. In future, it is possible to solve more real time problems in molecular biology.

References

1. Murugan, A., Lavanya, B.: A DNA algorithmic approach to solve GCS problem. Journal of Computational Intelligence in Bioinformatics 3(2), 239–247 (2010)
2. Murugan, A., Lavanya, B., Shyamala, K.: A novel programming approach for DNA computing. International Journal of Computational Intelligence Research 7(2), 199–209 (2011)
3. Delcher, A.L., Kasif, S., Feischmann, R.D., Peterson, J., White, O., Salzberg, S.L.: Alignment of whole genomes. Nucleic Acids Research 27, 2369–2376 (1999)
4. Lavanya, B., Murugan, A.: A DNA based approach to find closed repetitive gapped subsequence from a sequence database. International Journal of Computer Applications 29(5), 45–49 (2011)
5. Lavanya, B., Murugan, A.: Discovery of longest increasing subsequences and its variants using DNA operations. International Journal of Engineering and Technology 5(2), 1169–1177 (2013)
6. Lavanya, B., Murugan, A.: Multidimensional longest common subsequence discovery from large databases using DNA operations. International Journal of Engineering and Technology 5(2), 1153–1160 (2013)
7. Schensted, C.: Longest increasing and decreasing subsequences. Can. J. Math 13, 179–191 (1961)
8. Crochemore, M., Porat, E.: Fast computation of longest increasing subsequences and application. Information and Computation 208, 1054–1059 (2010)
9. Knuth, D.E.: Permutations,matricesand generalized young tableaux. Pacific. J. Math 34, 709–727 (1970)
10. Esmaieli, M., Tarafdar, M.: Sequential pattern mining from multidimensional sequence data in parallel. International Journal of Computer Theory and Engineering 2(5), 730–733 (2010)
11. Hirosawa, et al.: Comprehensive study on iterative algorithms of multiple sequence alignment. Computational Applications in Biosciences 11, 13–18 (1995)
12. De B. Robinson, G.: On representation of symmetric group. Am. J. Math 60, 745–760 (1938)
13. Fredman, M.L.: On computing the length of longest increasing subsequences. Discrete Mathematics 11(1), 29–35 (1975)
14. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. Journal of Computational Biology 1, 337–348 (1994)
15. Tompa, M.: An exact method for finding short motifs in sequences with application to ribisome binding site problem. In: Proc. Seventh Int'l. Conf. Intelligent Systems for Molecular Biology, pp. 262–271 (1999)
16. Bespamyatnikh, S., Segal, M.: Enumerating longest increasing subsequences and patience sorting. Information Processing Letters 76(1-2), 7–11 (2000)
17. Hunt, J.W., Szymanski, T.G.: A fast algorithm for computing longest common subsequences. Communications of ACM 20(5), 350–353 (1977)

A Peer-to-Peer Dynamic Multi-objective Particle Swarm Optimizer

Hrishikesh Dewan[1,2], Raksha B. Nayak[2], and V. Susheela Devi[1]

[1] Department of Computer Science & Automation
Indian Institute of Science, Bangalore
{hrishikesh.dewan,susheela}@csa.iisc.ernet.in
[2] Knowledge & Innovation
Siemens Corporate Technology & Development Center, Bangalore
raksha.nayak@siemens.com

Abstract. Multi-objective optimization problem is an important part in solving a wide number of engineering and scientific applications. To-date, most of the research has been conducted in solving static multi-objective problems where the decision variables and/or the objective functions do not change over a period of time. In a dynamic environment, the particles non dominated solution set during a specific iteration may no longer be valid due to change in the underlying system. As a result, traditional techniques for solving static multi-objective functions cannot be applied for solving dynamic multi-objective functions. Further, with the increase in the number of variables/objective functions, a single system based optimizer will take a long time to compute the non-dominated solution set. In this paper, we present a peer-to-peer distributed particle swarm optimization algorithm that tracks the change in the underlying system and is able to produce a diversified and dense non- dominated set using a network of peer-to-peer system. Our algorithms are tested using a set of known benchmark problems and results are reported. To our knowledge, this algorithm is the first of its kind in the areas of peer-to-peer particle swarm optimization.

1 Introduction

Dynamic multi-objective problems (DMOP) are an important class of optimization problems wherein the decision variables or the objective functions change over a period of time. The challenge in DMOP is to track the changing Pareto optimal front and produce non dominated (ND) results that are not only diversified but also close to the true Pareto optimal front. Formally, a DMOP can be defined as in 1. In 1, a number of minimization function objectives are defined. Each of these functions accept a vector x which is defined within a certain range. Further, there are J number of inequality constraints and K number of equality constraints. Each of these have a time parameter t which implies that based on discrete time steps these functions may change their respective function formulations. The formulation as noted in 1 is a constrained

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 453–465, 2013.
© Springer International Publishing Switzerland 2013

dynamic multi-objective function. In this paper, we focus on unconstrained dynamic multi-objective functions. This means we omit the inequality and equality constraints from the equation as given in listing 1. There are several researches that have been conducted in solving DMOP. [1], [2] and [3] are examples of particle swarm optimization (PSO) algorithms that have been used to solve such problems. All these algorithms work in a single computing system where the algorithm spans threads by taking advantage of the multiple cores as present in the system. However, with the increase in the number of variables and/or objective functions, such algorithms will take a large amount of time to compute the pareto front. One extension of these parallel threads is to use a dedicated cluster of machines. A dedicated cluster of machines with almost homogenous hardware/software and internetwork is, however, costly and requires dedicated maintenance.

On the other hand, peer-to-peer (P2P) distributed systems provide a viable alternative where individual machines can participate in a computationally rich optimization problem without any adherence to the strict administrative control as that of a cluster. A P2P system is characterized by a loose connection of independent machines using an overlay network where nodes can join and leave at their own wish. Being a loose cluster of machines, the design and use of a P2P network for optimization, however, has a number of challenges. First, in a P2P network there is no guarantee of node availability. A node may join, contribute for a brief period and exit altogether from the network. Further, even if a node is present, due to heterogeneity or variance in the configuration of hardware and software capability among the P2P nodes, there could be an unequal distribution of jobs across the system. Both of the above problems are crucial from the purview of a computationally rich job such as an optimization algorithm. The solution to the first problem of non-availability of nodes is important as without it an optimization problem will never be complete and the solution to the heterogeneity is important because without it, there would be unequal distribution of load. Unequal distribution of load may lead to inefficient utilization of the resources present in the P2P system. Thus a P2P system modeled for solving optimization problems does not only have to ensure availability of nodes for continuous optimization of the problem but should also have the provision of completely utilizing the resources of the system.

In this paper, we present such a distributed P2P system that aims in finding non-dominated (ND) Pareto optimal front for a dynamic multi-objective optimization problem. Our algorithm is based on particle swarm optimization and we have modified certain aspects of it for efficient detection of change in the environment and movement towards the changing Pareto optimal front. Our algorithm consists of a number of sub-swarms, where each node runs a single swarm optimization method. Each sub-swarms connects to other sub-swarms with a deterministic and a random method thereby facilitating both exploitation and exploration. We also handle fault tolerance of nodes in the system and distribute load evenly across the system. The algorithm has been tested with a wide range of available benchmark functions and in all cases we have achieved

results which are at par with the single computing system. The architecture of P2P network is elastic and it can scale well to thousands of P2P nodes without any significant reduction of performance. Although, there is a large body of work in the areas of parallel particle swarm optimization and such other swarm intelligence techniques, there is, however, no work that directly tries to optimize a dynamic multi-objective optimization problem in a P2P network.

$$\begin{array}{cccc} f_i\,(x,t)\,, & where & x=(x_1,x_2,x_3\cdots x_n) \\ g_j & (x,t)\leq 0, & j=1,2,3\cdots J \\ h_k & (x,t)=0, & k=1,2,3,\cdots K \\ x_i & = & [low_i, high_i] \end{array} \tag{1}$$

This paper is organized as follows. In section 2, we describe the related work and compare our work with the existing body of research. In section 3, we define the P2P network and the algorithms that are used to solve the dynamic multi-objective optimization problem. In section 4, we show the results of the experiments and analysis of the results and finally in section 5, we conclude the paper and provide the directions of future research in this area.

2 Related Work

There is not much work done on the DMOP and most of the work is based on algorithms that run in a single system. For example [1] is a modified particle swarm optimization technique that is based on an earlier particle swarm optimization variant [4]. In this paper, the authors divide the work of the Pareto front generation into distinct layers. The lower layer is responsible for optimization and the upper layer is used for maintaining the archive. This method is partly similar to our approach in that we also have sub-swarms for detection and optimization, but we do not have two different layers. Each sub-swarm is equal and they share ND set across their neighbors. Further, in our set up, the archive is maintained not in a single system but a group of replicated system and whenever a non dominated set is added, it is multicast to all the archives.

Further, there are numerous attempts to solve DMOP using several techniques such as evolutionary computing, memetic, immune system etc. All these algorithms [5], [6], [7]. [8], [9] and many more as described in [1] are all single system based optimizers and there is no P2P system used to distribute and solve the problem.

A close match to our work is [10]. However, as mentioned in [10], the author describes not a distributed but a parallel approach for solving dynamic optimization problems. In [10] , a master node is being defined and it is responsible for distributing work across a set of machines and for deciding the pareto front. As evident from the perspective of distributed system design, master nodes are the points of failures and contention. Hence, the system as defined in [10] does not scale up much and is ideal for a cluster of machines which are dedicated. Further extension and discussion of parallel evolutionary computing for DMOP is defined in [11] and [12] . However, all of these systems are based on a dedicated

parallel processing hardware such as high performance super-computing sites. In comparison, our system does not have any master nor requires a dedicated expensive high performance computing cluster. Every node is a master/slave at the same time and all message updates are asynchronous. Further, due to our redundant design, the notion of fault tolerance is almost ruled out. Hence, a direct comparison with the results that are obtained and of the technique used by them is not feasible.

3 Architecture and Algorithm

The entire distributed P2P framework for DMOP can be divided into two stages. The first stage is the network architecture and algorithms that are used for creating and maintaining the network. The second stage is the algorithm used for computing the ND front of the optimization problem. The first part is the middleware that is involved in keeping the distributed systems intact while the second part is the application itself that runs on top of this distributed system. However, since swarm intelligence requires information dissipation, the routing table of the underlying overlay network of the distributed system is partially populated and guided by the application at the top. In subsection A, we present the former whereas in subsection B we have presented the latter.

3.1 Network Architecture and Algorithms

We use an overlay network of nodes for communication of systems as well as application level messages. The overlay network is a circular ring and is formed by applying consistent hashing functions. When a node joins the network, it is provided with a unique 160 bit identifier. The identifier is derived by concatenating time and MAC address of the node as an input to SHA-1 hash function. Based on the output, the ring is organized lexicographically. Figure 1 is an illustration of the same. For routing messages from one node to the other, each node maintains a routing table. The routing table entries are of three different types. Lexicographic neighbors (bi-directional) in both the clockwise and anti clockwise directions (a configurable number each in either direction), second unidirectional random links (based on the number of 1s in the identifier) and third objective space neighbors. Objective space neighbors are the areas in function space which are adjacent to the node that is currently being computed. Each node at any single point of time computes only one quadrant. For 2D space the number of neighbors is 4 and in general for n-dimensional space there are 2n neighbors. The reason for maintaining the objective space neighbors is to refer to it during the strategic re-work negation algorithm defined in the subsequent sections. When a node needs to send a message to another node, the routing table for the node is consulted to find the nearest node. Nearest node distance is the lexicographic distance. The algorithm then forwards the message to the nearest node and the same procedure is repeated till either the destination node is reached or there is no node of that identifier. In case of the former, the message is consumed by

the destination node whereas in case of the latter, the message is either consumed by the most preceding node or an error is returned to the user. Due to non-uniformity of node links, the reverse path can be different from the source node. It can be proved that on an average the maximum number of hops required to transfer a message is not more than log N, where N is the number of nodes in the network. Also, due to non-uniformity of processing capability, each node may compute widely disparate function spaces. As such, therefore, the objective neighbor space changes from time to time and is not bound to be static during the entire life cycle of a node. Our network differs from other DHT-based networks such as Chord [13], Pastry [14] etc. by including extra neighbor node information related to our objective function. This modification is done to decrease the message overload and also for fast convergence. When a node completes its job of finding the Pareto front, it does not seek to enter into some other nodes dimensional space. Instead, the completed or the lightly loaded node simply skips the region of space and moves to the space which is slowly progressing or not yet processed. Thus, unlike the traditional PSO algorithms where there is a large amount of duplication of work, this design ensures very little duplication of work. Hence, compared to other algorithms, the relative time to convergence is faster.

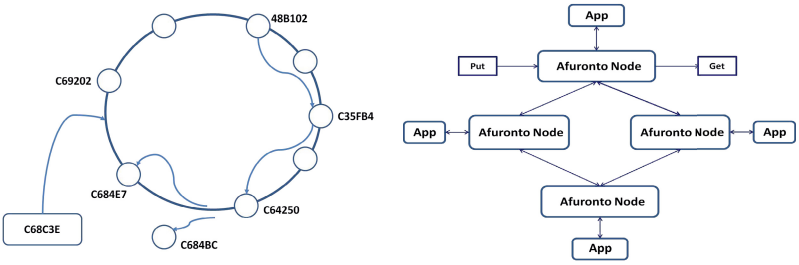


Fig. 1-2. An overlay network of nodes. and Peer-Peer PSO Architecture

However, due to large number of dimensions, the effective neighbors are roughly around 2N. This is a large number even for moderately sized dimensions. To reduce the effect of introducing a large number of such neighbor nodes, we include a tree based directory structure . Every node maintains a single indirection to its parent and each parent includes the list of siblings. The number of siblings is not more than the number of dimensions that is being used to divide the hyper-plane. Thus, the maximum number of neighbors that a node has to store is dependent upon the problem statement and also on the factors of division. For example, for three dimensional vectors, we first divide the function space in one of its axis. The second level includes the divisions of the second axis and so on. This increases the time to propagation but it prohibits the overcrowding of the routing table entries of the node.

3.2 Node Management

In a peer-peer network all nodes are symmetric and equal. Whenever a node has to participate in the function optimization computation, the first operation

is the join operation. Upon completion of join, a node acquires a membership to the network and participates in all its activities, which includes function optimization, message passing, load balancing and fault-tolerance. Since each peer-node is responsible for atleast a certain chunk of work, there must be well-defined node leaving protocol. In this section, we define the node joining and leave protocols. It is important to note that in an ad-hoc network with no central administrative unit, nodes can join and leave at any moment. Such abrupt change in the network topology could be due to malfunctioning of hardware, software or the network. Hence, the join and the leave protocol must handle these extreme but major use cases as well.

After a new node is created, it searches for an existing node in the network to communicate with. The list of existing nodes must be supplied to the new node before it starts its operation and is, by and large, a manual process. Once the node is connected to any of the existing nodes, the first message it composes is to find duplication of its own calculated node ID. Such a message is routed to the network using the newly connected existing node. If there are no such nodes, then the node can safely join the network. On the other hand, if such a node exists, the new node creates a new ID and recursively follows the same process. Joining the network involves creating the routing table entries. After the routing table entries are successfully created, the node connects to any node in the network for work units. Since a node already existent in the network is completely participating in the optimization process, the new node requests work from the already existent nodes. In our present protocol, we allow a new node to acquire work from a number of existing nodes and select the work that seems appropriate. Upon selection of the work unit, the neighbors nodes are identified and the objective neighbor node work is filled. At this stage the routing table entries are complete and also the necessary code and work unit is with the new node. The new node starts executing its own sequential PSO from this point. Note that this type of work selection is completely deterministic and may take some time if the already existing nodes do not have work to allocate to a new node. The other joining protocol that we have investigated is to randomly select a work. Once the new node acquires the necessary code for the functions, its solution space and work division technique, it randomly selects a work unit from the available work units. Since the routing table entries maintain neighbor node information based on the decision space neighbors, a message is sent from the new node to the node which is responsible for the work chunk. If the work chunk is available, then it is allocated to the new node or the node recursively continues the random work unit selection process until it gets the desired work load. It is important to note that if the number of nodes are larger than the available work units, then neighbor nodes can further partition its work space and allocate a few to the newly joined node.

The node leaving protocol is the opposite of what is defined in the node joining protocol. The leaving node broadcast “leave messages to all its neighbors and waits for a few seconds before it leaves the network. Upon leaving the network, a “hole” is created in the overlay network and routing table entries of the neighbor

nodes are modified. Dynamic node churns are described in more detail in a later section.

3.3 Information Propagation

Information Propagation is done using the routing table entries of each node. It can hence be seen that each node in the network works as a router and is responsible for forwarding messages to the closest (lexicographically) destination node for transmission. In our system, there are two types of information. The first type is the information related to swarm optimization and second type is the information for maintaining the network. Since in a multi-objective PSO, there is no individual gbest but a set of ND solutions, the only information that is shared across peer swarm agents is the ND set. However, unlike gbest transfer in a PSO, we need not transfer the set of ND solutions to all of the nodes. As noted earlier, our objective space is divided into distinct hyper-planes and we maintain a hierarchy of neighbors. The ND solution is only passed to the neighbors. The receiving node takes the responsibility of finding dominated solutions, if any, from this set. The other information that is passed in the network is related to the maintenance of the network. Node join and leave, heartbeat messages and that of load balancing information are some of them. More details of these messages are mentioned in their respective subsections.

3.4 Load Balancing and Fault Tolerance

A peer-peer ad-hoc network is always a mix of diversified components: diversified in terms of hardware resources and software components available for computation. Therefore, there is no uniformity in the completion time of a solution space. Some nodes may take a long time to compute a work unit whereas some nodes may complete the same in 1/10th of the time taken by the other node. As a result, load balancing of nodes is an important requirement in such a diversified peer-peer network. We balance loads not instantaneously, but after repeated step intervals. A step interval is a finite number of iterations. After completion of each step, the node propagates its load to the neighbors using a broadcast. Nodes that are lagging behind comparatively are further propagated.

For broadcasting, there are two specific rules. For each node we maintain a least and utmost load, which are respectively 20% and 80% of its load capacity. If the CPU utilization falls below or above this limit, the node broadcasts this information to the neighbors. Every node therefore maintains the load of its neighbors. If, however, the load is not below or above this threshold limit, there is no message sent. Hence, load information table is not as populated as the routing table entries. Once a node receives such information, it compares the load with all the entries and tries to achieve equilibrium by matching low capacity nodes with the high capacity ones. If, on the other hand, there are nodes that are still not yet matched, then the information is passed on to the nodes neighbors. The process is repeated until either there is a match or there are no nodes to match. When a node completes its allocated work unit and there are no more pending

works, the computational utilization decreases by 10%. Under this circumstance, the broadcast is sent from the node to all its neighbors till it receives new chunk of data. It is not difficult to prove that lowest utilized node is always eventually matched with a highly loaded node and the maximum number of steps required to match such a node is no more than log N. Thus the network tries to balance load at the neighborhood first and if unsuccessful, propagates the information to the next level. With this, there are no central co-coordinators required and also the number of messages required for balancing the load is less.

As in the case of load balancing, fault tolerance is also handled co-operatively. Each node upon joining the network maintains three fault tolerant connections to three other nodes. These nodes need not be entries in the routing table and are independent of them. After every successful time interval, which is configurable, the node sends the best positions and information on work units in allocation to each of these nodes. Failure to receive updates by the majority, either due to its software/hardware or network partition, signals the node as dead and a new node is selected for execution of the work.

```
Algorithm : Particle Execute
Description : This method is executed by each particle in the PSO
Input : A randomly chosen gbest,bool value
Result: Null or particle's own best position
Begin

changeDetected by this particle ← false;
if (currentPosition is a part of global non − dominated set)
pBest=currentPosition;
if (iteration count % curve check factor)==0 then
    if (best result of this particle != ComputeFunction(pbest of this particle) then
        best result of this particle ← null;
        pbest of this particle ←null;
        changedetected by this particle ← true;
    end
end
if changeDetected by this particle ←false then
    update velocity;
    update current position;
end
result ← ComputeFunction(n-dimensional currentPosition);
if (result dominates bestresult) then
    pbest of this particle ← current position;
    bestResult of this particle ← result;
    pBest ← pbest of this particle;
    result ← bestResult of this particle;
    changeDetected ← change detected by this particle;
end
else  return null;
return (pBest,result,changeDetected);
End
```

Algorithm 1. Execution of a Particle

3.5 Optimization Algorithm

Our optimization algorithm is designed to handle all possible cases of change detection in the environment. In a dynamic environment, when there is no access to the optimization functions, change may happen in multiple points in the valid search landscape. For example, change could be only minimal resulting in a slight change in the theoretical Pareto optimal front or there could be a complete change in the Pareto front (for example, from a convex to a concave). Further, changes may happen in two different ways-either there is a change in the decision variable space or there is a change in the optimization in the objective space.

If the type of change is known apriori, it is relatively trivial to reposition the particles in the changed landscape and find solutions close to the true Pareto optimal front. Since the change in the environment cannot be explicitly determine, a range of pattern recognition techniques together with time series analysis can be used to predict the change. However, such predictions will require historical data and we refrain from discussing it in this paper. In the future, we would like to integrate such prediction techniques for better positioning of particles in the changing search landscape.

```
Algorithm : Particle Swarm Optimization
Input : Input dataset
Result: Optimized position
Begin
Initialize N particles including old type and new type particles;
Initialize Gbest to null;
while Iteration count < Total number of iterations do
    for p ← 0 to total number of particles do
        (pBest,result,changeDetected ← ExecuteParticle;
        if result ! = null then
            if changeDetected==true then
                gbest ← null;
                Notify other particles about the change;
                Re randomize T% of the old type particles;
            end
            else
            if result is one among the non − dominated set elements then
                Add result to the global non − dominated elements set;
            end
            mother particle ← a randomly chosen gbest particle among the non − dominated
            set;
            if (mother particle ! = new born particle type) then
                Call Give Birth to new born particles() of mother particle;
                Add the new born particle list to the main list;
                Delete the old new born particle list from the main list;
                if (mother particle == new particle type) then
                    Re-randomize T% of the old type particles;
                end
            end
        end
    end
    Increment the Iteration count;
end
Non − dominated set is the result;
End
```

Algorithm 2. PSO Algotithm

In our present algorithm, we have two different types of particles namely the new particle types and the old particle types. New particle types are like foragers and they wander in the solution space within a fixed radius. Old particles are just like the old particle swarm particles and they follow the leader which has the best values. Every particle stores the location where it achieved the ND point in its memory and after a pre-defined step,it recomputes the ND set once again. This is being done to check if there is a change in the environment. If there is a change in the environment, the change notification is sent to each of its neighbors by a broadcast. The change notifications are high priority messages and they are broadcast recursively. Since change in the environment can happen at multiple different search spaces and many particles detect it, the number of broadcast messages could be enormously high. To reduce this, we use a threshold. Beyond a certain threshold within a fixed interval, if a particle receives change notifications, further notifications are simply suppressed. Therefore, such change notifications only generate a small number of messages. Further, if a particle received a change notification, and thereafter recognizes a change in its landscape within the pre-determined interval, the change notification is also suppressed.

The new born particles do not have any memory and they just report the ND values.

In the distributed system, each process executes a subswarm and the number of particles, both old and new, can vary. Such variations are necessary as machines can have varying configurations. Further as noted in section 3.1, each subswarm is only responsible for its search landscape. The subswarm's exchange messages among themselves and also with a set of repositories which is finally responsible for calculating the final pareto optimal front.

4 Experiment and Results

The P2P framework together with the modified particle swarm optimization as mentioned in section 3 has been tested with a large variety of benchmark optimization problems. The code for the particle swarm optimization and P2P network is written using C# and .NET 4.0. For simulation of the algorithm, we created a ten node cluster where each node runs a number of peer processes. Each node in the cluster is heterogeneous in both software and hardware. Each process is a P2P node and encapsulates all the protocols etc as mentioned in section 2. A P2P node in our simulation holds at least 10 old particles and 5 new particles but in the actual run, the number of such particles can be configured and extended to as many as desired by the user. For the simulation of uneven load distribution, the parameters alpha and beta are initialized to different values. Further, we maintain a separate control process that arbitrarily creates new processes and kills existing processes in each single system for simulating node join and leave. For testing the system, we have used a number of benchmark problems. In this paper, we show the results for three such popular multi-objective benchmark functions, FDA1, FDA2 and FDA3 as mentioned in [15]. Other benchmark functions as proposed in [16][17][2][3] are also tested and evaluated. Besides these, a number of other functions that were tested with this algorithm can be found in [1]. However, due to lack of space, we have omitted the results of such functions. In table 3,4 & 5, we have shown the results of the computation of FDA's with the well known metrics. The metrics are defined in [1] and in this paper we have used average error.

In figure 3-8 we have shown the convergence graph for each of these particles. In FDA 1, the theoretical pareto front does not change and hence it remains constant. The other plots in the same graph show the results of convergence after successive iterations. For each function, we have shown results with FES/particles as 25,000/50 and 1,00,000/100. Each experiment is executed for twenty five runs and the results of the twenty five runs are tabulated in Table 3,4 & 5. Further, each function changes its internal representation after every five iterations. Apart from these results, we have also experimented with the effect of convergence graphs for node churns and stimulated increased load. The results for the node churns are not included in this paper for lack of space.

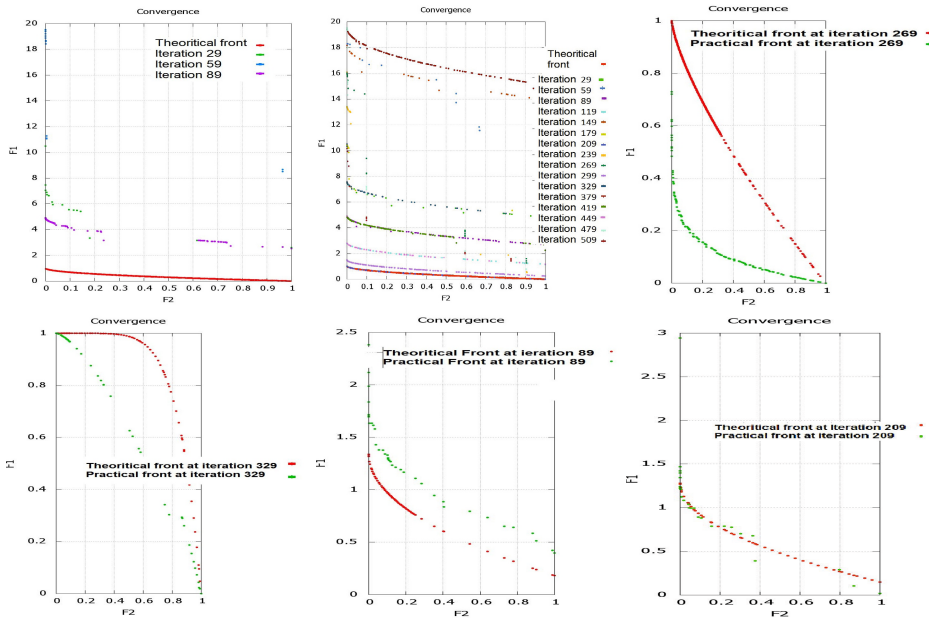


Fig. 3-8. illustrating FDA1, FDA 2 & FDA3; Figure 3. FDA1 FES 125,000, Particles-250 , Figure-4 FDA 1 FES 150,000, Particles- 100; Figure 5 FDA2 FES 2,50,000, Particles-100, Figure -6 FDA2 FES 2,50,000, Particles- 100; Figure 7 FDA3 FES 100,000, Particles-250, Figure -8 FDA3 FES 300,000, Particles- 250.

Table 1. Tabulated Results of FDA1, FES- 250,000, Particles : 50, Function changes every 5 iterations

Metrics	S	ER	GD	CM	HV	Spread	Max Spread	SC	WM
Best	0.025	0.201	0.024	0.085	4.246	0.801	1.574	0.208	0.577
Worst	1.436	1.000	4.565	16.763	19.589	1.182	18.930	1.000	2.065
Average	0.328	0.947	1.712	7.049	13.687	0.999	6.964	0.950	1.213
Variance	0.178	0.040	2.045	29.142	22.618	0.010	24.005	0.039	0.204
Std dev	0.374	0.193	1.419	5.394	4.753	0.095	4.873	0.192	0.448

Table 2. Tabulated Results of FDA2 , FES- 250,000, Particles : 50, Function changes every 5 iterations

Metrics	S	ER	GD	CM	HV	Spread	Max Spread	SC	WM
Best	0.008	0.000	0.004	0.012	18.999	0.659	1.147	0.000	0.474
Worst	0.349	0.901	0.060	0.315	19.908	1.352	2.374	0.726	0.951
Average	0.053	0.596	0.029	0.206	19.647	0.880	1.394	0.048	0.625
Variance	0.013	0.104	30.0E-05	0.011	0.098	0.033	0.434	0.033	0.016
Std dev	0.090	0.323	0.015	0.107	0.313	0.180	0.342	0.176	0.124

Table 3. Tabulated Results of FDA 3 , FES- 250,000, Particles : 50, Function changes every 5 iterations; Abbreviation for metrics : S - Spacing, ER - Error Ration, GD- Generational Distance, CM- Convergence Metric, HV- Hypervolume, SC- Set Coverage, WM- Weighted Metrics

Metrics	S	ER	GD	CM	HV	Spread	Max Spread	SC	WM
Best	0.053	0.235	0.060	0.146	1.832	0.586	0.861	0.397	0.446
Worst	0.532	0.993	0.324	0.857	19.329	1.144	3.768	1.000	0.856
Average	0.196	0.655	0.147	0.408	14.699	0.873	2.245	0.838	0.655
Variance	0.019	0.052	0.006	0.034	37.108	0.025	0.677	0.035	0.013
Std dev	0.133	0.227	0.071	0.181	6.090	0.155	0.816	0.184	0.114

5 Conclusion

Investigation in distributed swarm intelligence algorithms for optimization is an important necessity for solving complex optimization problems of today. The current problems which are inter-disciplinary in nature, may involve hundreds of variables and objective functions. Further, due to a continuous change of environmental parameters, which is generally the case with a number of real life optimization problems, such optimization frameworks hold the answer for solving tomorrow's problem. In this paper, we describe the framework and just a handful of results. We omitted sections on dependability and security designs of the P2P network. In a P2P network such features are essential. A P2P network is trust deficient after all and malicious nodes can not only subvert messages but also deceptively inform wrong values. In the future, we would have a more elaborate publication that will include features such as these and more results that exhibit the elasticity and load distribution of the network. Further, we believe that dynamic optimization problems in some cases, exhibits patterns. Such patterns are temporal in nature and hence the present framework may be extended to include pattern recognition and machine learning techniques for prediction of change in the environment and re-positioning of the particles in the search landscape.

References

1. Helbig, M.: Solving dynamic multi-objective optimisation problems using vector evaluated particle swarm optimisation. PhD thesis, University of Pretoria (2012)
2. Wang, Y., Li, B.: Investigation of memory-based multi-objective optimization evolutionary algorithm in dynamic environment. In: IEEE Congress on Evolutionary Computation, CEC 2009, pp. 630–637. IEEE (2009)
3. Tang, M., Huang, Z., Chen, G.: The construction of dynamic multi-objective optimization test functions. In: Kang, L., Liu, Y., Zeng, S. (eds.) ISICA 2007. LNCS, vol. 4683, pp. 72–79. Springer, Heidelberg (2007)
4. Schaffer, J.D.: Multiple objective optimization with vector evaluated genetic algorithms. In: Proceedings of the 1st International Conference on Genetic Algorithms, pp. 93–100. L. Erlbaum Associates Inc. (1985)

5. Wang, Y., Dang, C.: An evolutionary algorithm for dynamic multi-objective optimization. Applied Mathematics and Computation 205(1), 6–18 (2008)
6. Shang, R., Jiao, L., Gong, M., Lu, B.: Clonal selection algorithm for dynamic multiobjective optimization. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-m., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 846–851. Springer, Heidelberg (2005)
7. Zeng, S.Y., Chen, G., Zheng, L., Shi, H., de Garis, H., Ding, L., Kang, L.: A dynamic multi-objective evolutionary algorithm based on an orthogonal design. In: IEEE Congress on Evolutionary Computation, CEC 2006, pp. 573–580. IEEE (2006)
8. Moscato, P.: On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Caltech concurrent computation program, C3P Report 826 (1989)
9. Zhang, Z., Qian, S.: Artificial immune system in dynamic environments solving time-varying non-linear constrained multi-objective problems. Soft Computing 15(7), 1333–1349 (2011)
10. Cámara, M., Ortega, J., Toro, F.J.: Parallel processing for multi-objective optimization in dynamic environments. In: IEEE International Parallel and Distributed Processing Symposium, IPDPS 2007, pp. 1–8. IEEE (2007)
11. Cámara, M., Ortega, J., de Toro, F.: Approaching dynamic multi-objective optimization problems by using parallel evolutionary algorithms. In: Coello Coello, C.A., Dhaenens, C., Jourdan, L. (eds.) Advances in Multi-Objective Nature Inspired Computing. SCI, vol. 272, pp. 63–86. Springer, Heidelberg (2010)
12. Ruiz, I.R.: Sinta-cc: Adaptive intelligent systems for modelling, prediction and dynamic optimization in clusters of computers tin2004-01419
13. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. ACM SIGCOMM Computer Communication Review 31, 149–160 (2001)
14. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In: Guerraoui, R. (ed.) Middleware 2001. LNCS, vol. 2218, pp. 329–350. Springer, Heidelberg (2001)
15. Farina, M., Deb, K., Amato, P.: Dynamic multiobjective optimization problems: test cases, approximations, and applications. IEEE Transactions on Evolutionary Computation 8(5), 425–442 (2004)
16. Koo, W.T., Goh, C.K., Tan, K.C.: A predictive gradient strategy for multiobjective evolutionary algorithms in a fast changing environment. Memetic Computing 2(2), 87–110 (2010)
17. Goh, C.-K., Tan, K.C.: A competitive-cooperative coevolutionary paradigm for dynamic multiobjective optimization. IEEE Transactions on Evolutionary Computation 13(1), 103–127 (2009)

Reduce Energy Consumption through Virtual Machine Placement in Cloud Data Centre

Nongmaithem Ajith Singh and M. Hemalatha

Department of Computer Science, Karpagam University,
Coimbatore, Tamil Nadu, India
{ajithex,csresearchhema}@gmail.com

Abstract. In this paper, energy consumption in the data centre was studied where thousands of servers and other devices runs, energy are utilized to run the server and cooling the environment. Energy consumption can be reduced by switching off the idle server by means of migration of Virtual Machine from under-load Host. Load in cloud computing is maintained by migration of VM from the overloaded Host to a free Host or activate new Host. Based on this study, a reservation technique by using BIN packing was proposed in this paper with an overload detection algorithm. The proposed algorithm RBIN is experimented in 800 servers with 1024 Virtual Machines. From the experimental result, proposed method RBIN reduces energy in higher level.

Keywords: Cloud Computing, energy, virtual machine, RBIN, PR.

1 Introduction

The long-term vision of computing is to provide as a service to the public by subscribing the technology. The evolution of computer hardware, internet, programming language , various applications and gadget has able to use the computing as a service which is termed as cloud computing, where any person around the globe can utilize computing by using any device which has an internet connection and an internet browser. Cloud computing defined by Gartner is "A style of computing where massively scalable IT enable capabilities are delivered as a service to external customers using internal technologies"[1].

Cloud computing is a service rather than a product and service in the sense its computing resources. Cloud 'Services' refer to those types of services provide by the cloud provider (CP) and that can be used by cloud customers on a 'pay-per-use' basis. Cloud services are classified as Infrastructure as a Service (IaaS), Software as a Service (SaaS), Platform as a Service (PaaS) [2]. End user can subscribe to any services mention above as much or as little computing power, as they need and can pay per use, which is infinite in nature, user can start the service or stop the service. From a CP perspective, the key issue is to maximize profits by minimizing the operational costs [3]. The Cloud provider goal is to provide services, which the user request in an efficient manner without any violation of the SLA.

Success of cloud computing is the mutation of many technology like grid computing, cluster computing, SOA, Virtualization. Among these, Virtualization is

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 466–474, 2013.
© Springer International Publishing Switzerland 2013

the most important. Virtualization is a technology originally developed for mainframe computing, in which by using hypervisor on above the bare metal it could share the hardware resources which are CPU, RAM, Disk Space, each OS is run with different application and user are not aware that their resource is shared with others [3].

Thou the resources is shared it is fully under the control of the user. The benefit of using Virtualization is if there is a need of scaling the hardware resources it can be scaled up and if requires it can be moved from one host (Physical Machine) to another which is known as VM migration [4]. VM migration was done either to balance the load of the Host in a data centre or for saving the energy by switching off the idle host. Most Popular modern Virtualization technology products are VMware, Xen & KVM.

Example: Consider a data centre (DC) with five VMs, Z, X, B, H, P with different requirements and services running on it. Fig 1 represent 5 different services running on different server, Figure 2 represents services running on a VM on the same server where it uses 3 servers instead of 5 and figure 3 use 2 servers to run 5 VMs, by using cloud computing use of the servers is minimized.

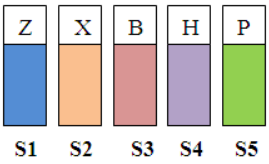


Fig. 1. DC without Virtualization

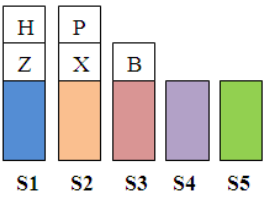


Fig. 2. Cloud DC with Virtualization

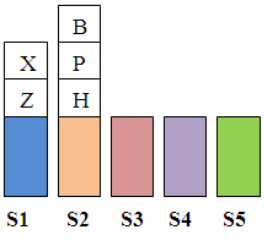


Fig. 3. Cloud DC with Virtualization

The energy cost to operate in the data centre is huge. The user has to pay for energy consumption apart for utilization of servers. Data centres consume a large amount of power that can consume a maximum of 50MW of power, which is equivalent to a city with 40K households. The cost can exceed to $15M per year.

2 Problem Description

Virtual machine placement is an NP hard problem different methodology had been proposed many researchers. The problem of virtual machine placement in the data centre is defined as: given a set of virtual machines $VM = \{vm_1, vm_2, ..., vm_n\}$ and a set of physical machines $PM = \{pm_1, pm_2, ...,pm_m\}$, where each vmi is a triplet $vm_i = (cpu_i, ram_i, bw_i)$, $1 \leq i \leq n$ denoted cpu, memory and bandwidth requirements of virtual machine respectively. Each pmj is also a triplet $pm_j = (cpu_j, ram_j, bw_j)$, $1 \leq j \leq m$ denoted resource capacity of the physical machine. In addition, x_{ij} , $1 \leq i \leq m$, $1 \leq j \leq n$ and y_i, $1 \leq i \leq m$ are decision variables, $x_{ij} = 1$ if and only if vm_j is mapped onto pm_i, $y_i = 1$ if pm_i is used to host virtual machine. The objective is to minimize

$$\sum_{i=1}^{m} yi \tag{1}$$

while finding all values of x_{ij}.

There are several implicit constraints in the above definition:

- Each virtual machine can only be hosted on one physical machine.
- For each type of resource, the amounts of resource requests of virtual machines sharing the same physical machine are smaller or equal to the capacity of the physical machine hosting them.
- The numbers of physical machines that host virtual machines are not more than [5].

$$m, \sum_{j=1}^{m} yi \leq m \tag{2}$$

To give a solution to this problem we proposed a method to reserve the resource while doing VM placement.

3 Related Work

Shu-Ching Wang et al.(2010) proffered a three-level cloud computing network highlighting the changes of network bandwidth and hardware technology and uses of the low-power host to achieve higher s reliability. The authors used OLB scheduling algorithm for balancing the load by keeping the system busy. The algorithm contrivances in such a way that all the host will be working state apart from that LBMM scheduling algorithm is also used to achieve the minimum execution time. This work was mainly focused on task scheduling by considering a dynamic nature of network in the cloud [6].

Rajkamal and Pushpendra (2012) proposed a rule based resource manager by highlighting the resource provisioning in a private cloud. The authors described that when more resources are required in private cloud it can be extended with a rule based method and can be used in the public cloud [7]. N. Bobroff *et al.* 2007 proposed an algorithm which allocates minimum resources to VMs and because of which it was able to predict the future resource requirement. Hence, their algorithm remaps the VM to PM for future resource demand [8].

Verma *et al*. 2009 proposed a 90th percentile based provisioning approach where they presented two algorithms, (i) Correlation Based Placement (CBP) in which each VM is sized at the 90^{th} percentile of its peak resource demand. For each VM to placed, it checked first whether it has a positive correlation with any of the VMs that placed in a particular machine. (ii) Peak Clustering Based Placement (PCP): with PCP approaches, each VM is provisioned with the 90th percentile utilization value, and a peak buffer of capacity equal to the maximum of peak size of all the VMs with considerably low correlation among their peaks of resource demand is kept reserved for all those VMs. It was observed that CBP has a palpable disadvantage of ending up using many servers when there are many correlated VMs or applications whereas PCP fixed this problem but still much scope for resource wastage was left due to the provisioning of resource for each VM individually and presence of a peak buffer [9]. Live migration of Virtual machine which is consider and support only on cloud

computing is deeply studied by Clark et al, live migration of virtual machine happens when a host is overloaded and to balance it live migration happen, author use Xen hypervisor for the experiment and proof that downtime of service is below discernable thresholds [10].

4 Allocation Policies

In general, the problem of dynamic VM consolidation can be split into 4 sub-problems:

a) Host under load detection.
b) Host overload detection.
c) VM selection.
d) VM placement [4] [11].

5 Proposed Method

The proposed method improved the BIN is packed with reservation technique and term as RBIN (Reservation BIN).

5.1 Reserve Modified Bin Packing

An approach which can taken to solve the VM placement is by using Bin Packing method [12], the physical machine can be considered as bins and the VMs to be placed can be considered as objects to be filled in the bin. Many researchers have solved the problem of VM placement by using Bin Packing. In the modification while allocating the VM to the server, the server reserved some resource for scalability. Two upper thresholds were being used on this proposal.

5.2 RBIN Pseudocode

Placement of VMs to a single host was done by keeping the host reserve with 30% of total resources; Allocation of VMs to Host is done by using BIN Method. Once the host overloads it, scale using reserve resource until it reach 90% to avoid migration

Step 1: Allocate VM if size < 70%
Step 2: If overload detection [**Step 4**]==true allocate 100-30%
Step 3: Else allocation size<90
Step 4: Migrate VM with selection policy [**Step 6**]
Step 5: Applies Overload Detection [Polynomial Regression]
Step 6: VM Selection [Available on CloudSim]
Step 7: Repeat Step 1

5.3 Overloading Detection - Polynomial Regression Models

- In general, we can model the expected value of y as an n^{th} order polynomial, yielding the polynomial regression model:

$$y = a_0 + a_1 x + a_2 x^2 + \ldots\ldots + a_n x^n + \varepsilon \quad (3)$$

- "Resource scheduler" data table named Vmtable consisting of the variables CPU , Memory , Costs, and Power.
- To run a polynomial regression model on one or more predictor variables, it is desirable to initial centre the variables by subtracting the matching mean of every, so as to ease the inter correlation among the variables [13].
- Suppose we wish to use a second order polynomial model involving the response variable CPU and the predictor variables Memory and Costs. To centre them, the R commands would be:
  - x1 <- VM$ Memory – mean(VM$ Memory) (4)
  - x2 <- VM$Costs – mean(VM$Costs) (5)
- Note that we have named the centred variables x1 and x2. We also will need the second order terms for the model:
  - x1sq <- x1^2 (6)
  - x2sq <- x2^2 (7)
  - x1x2 <- x1 * x2 (8)
- The names chosen are, of course, arbitrary. Obviously we could continue with a third order terms and so forth as needed.
- Next, we need to add these new variables to our data table:
  - VM <-(VM, x1, x2, x1sq, x2sq, x1x2) (9)

    Then we can obtain a second order regression model named Poly for these three variables in the usual manner:
  - VM <- CPU (x1 + x2 + x1sq + x2sq + x1x2) (10)
- x3 <- VM$ Power – mean(VM$ Power)
  - VM <- Power (x1 + x2 + x1sq + x2sq + x1x2) (11)
- if VM<- COST > Threshold
  - then Overload = True.
  - else Overload =false

6 Result and Discussion

The algorithm proposed RBIN in this work reserve the resources in cloud computing using of various VM placement techniques can reduce the VM migration, SLA and Energy. The proposed method first reserve the resource while VM allocation by 30% as reserve resource on the host, when the host reach 70% of its total resource it allows the VM to use the free resource of 30% with a threshold pointing at 90%. When the host reaches 90%, which can be scaled up or down, it, triggers the overload detection algorithm Polynomial Regression or policy available with the CloudSim.

When the host is overloaded, VM selection policy is activated which is available in CloudSim, which migrate the VM to another free Host or activate a new Host. Using PlanetLab Workload [4] simulation is run for 8 hours with proposed VM placement, Overload detection and VM selection Table 1 showing the experimental result of proposed VM placement RBIN runs with various overload detection and VM selection of Maximum Correlation, table 2 with a VM selection of Minimum Migration Time, table 3 with a VM selection of Minimum Utilization and table 4 with a VM selection Random selection based the experimental result on table graph chart is represented on figure 4 based on energy consumption, figure 5 for SLA violation and figure 6 for VM migration.

From the simulation, the proposed algorithm **RBIN** with **PR** and **MMT** gives the best result for **Energy with consumption of 23.83kWh** and for **SLA** overload detection of (**MAD, THR**) and VM selection (**MMT, MU**) gives the minimal SLA violation of **0.00031%**, for VM migration overload detection of **MAD** and VM selection **MMT** gives the minimal migration of **777 VM**.

Table 1. Maximum Correlation with Overload Detection

Overload Detection/VM Selection	ENERGY kWh	SLA in %	MIGRATION
IQR-MC	25.18	0.00033	840
LR-MC	24.68	0.00038	919
LRR-MC	25.51	0.00034	828
MAD-MC	25.28	0.00036	840
PR-MC	24.68	0.00038	882
THR-MC	25.73	0.00033	807

Table 2. Minimum Migration Time with Overload Detection

Overload Detection/VM Selection	ENERGY kWh	SLA in %	MIGRATION
IQR-MMT	24.28	0.00034	847
LR-MMT	25.65	0.00037	844
LRR-MMT	25.19	0.00034	872
MAD-MMT	26.56	**0.00031**	**777**
PR-MMT	**23.83**	0.00038	833
THR-MMT	25.36	0.00035	821

Table 3. Minimum utilization with Overload Detection

Overload Detection/VM Selection	ENERGY kWh	SLA in %	MIGRATION
IQR-MU	24.83	0.00034	842
LR-MU	24.84	0.00035	814
LRR-MU	24.07	0.00037	820
MAD-MU	27.15	0.00033	786
PR-MU	25.6	0.00032	832
THR-MU	25.94	**0.00031**	810

Table 4. Random Selection with Overload Detection

Overload Detection/VM Selection	ENERGY kWh	SLA in %	MIGRATION
IQR-RS	25.37	0.00035	854
LR-RS	25.29	0.00037	811
LRR-RS	24.78	0.00036	842
MAD-RS	24.37	0.00035	840
PR-RS	25.9	0.00035	859
THR-RS	25.89	0.00032	825

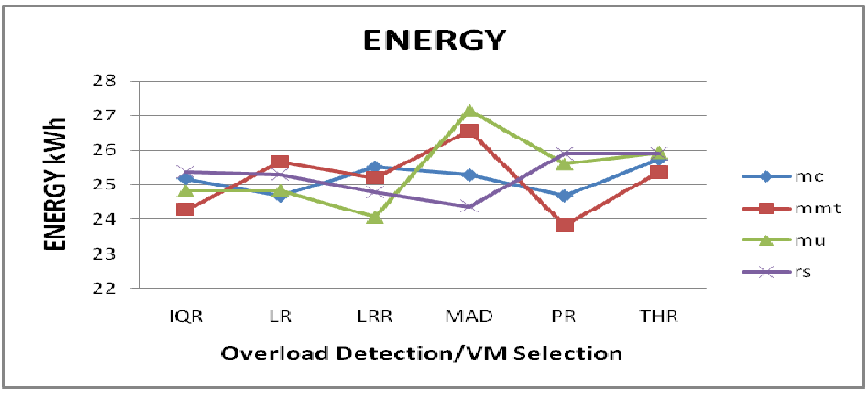


Fig. 4. Energy

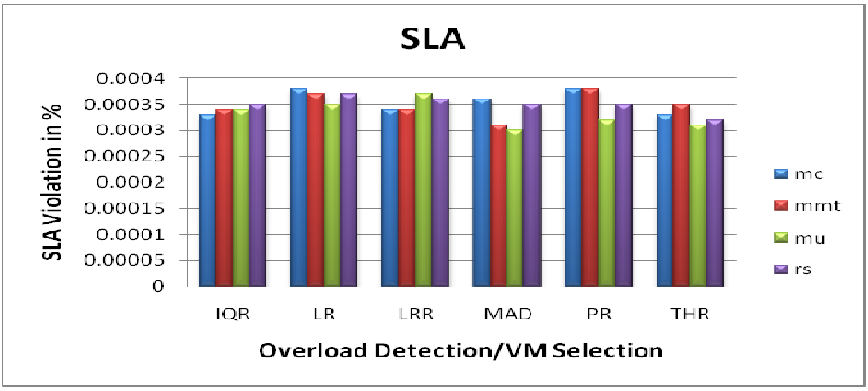


Fig. 5. SLA

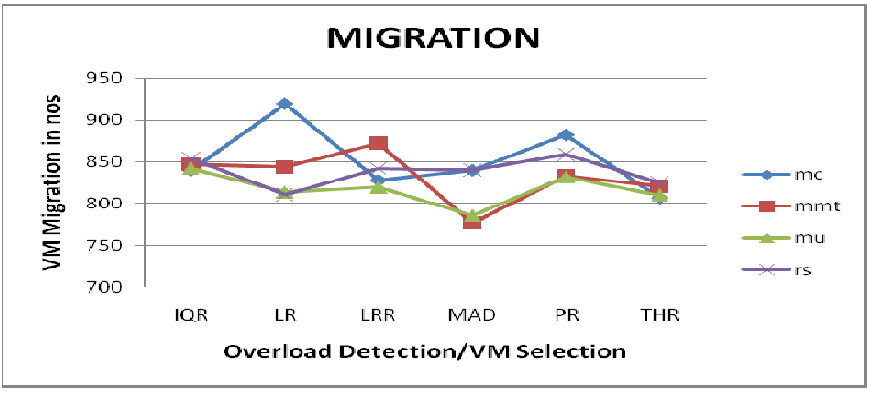


Fig. 6. VM Migration

Abbreviation

Overload Detection: DVFS-Dynamic Voltage Frequency Scaling, IQR-Interquartile Range, LR-Local Regression, LRR-Robust Local Regression, MAD-Median Absolute Deviation, THR- CPU utilization threshold, PR - Polynomial Regression, **VM Selection Policy:** MC-Maximum Correlation, MMT-Minimum Migration Time, RS-Random Selection, MUR - Minimum utilization Rank

7 Conclusion

The proposed method RBIN reduced the energy consumption in data centres, the idea of implementing the reservation technique with BIN balance the load, improved the VM placement, and reduces the energy. Proposed Overload detection of polynomial regression enchanted the RBIN. RBIN is simulated and compare with existing overload detection and VM selection. Virtual machine presents a great opportunity for cloud. Cloud provider has to consider minimizing the cost and the factor related to that is processor, storage, memory, network, and power.

References

1. Mell, P., Grance, T.: NIST. The NIST Definition of Cloud Computing (ver. 15). National Institute of Standards and Technology, Information Technology Laboratory (October 7 2009)
2. Randles, M., Lamb, D., Taleb-Bendiab, A.: A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing. In: IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, pp. 551–556 (2010)
3. Abdul-Rahman, O., Munetomo, M., Akama, K.: Live Migration-based Resource Managers for Virtualized Environments: ASurvey. In: The First International Conference on Cloud Computing, GRIDs, and Virtualization, pp. 32–40 (2010)
4. Esnault, A.: Energy-Aware Distributed Ant Colony Based Virtual Machine Consolidation in IaaS Clouds, http://dumas.ccsd.cnrs.fr/dumas-00725215/
5. Beloglazov, A., Buyya, R.: Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers. Software: Practice and Experience 41(1), 23–50 (2011)
6. Wang, S.-C., Yan, K.-Q., Liao, W.-P., Wang, S.-S.: Load Balancing in Three-Level Cloud Computing Network, 978-1-4244-5540-9/10/$26.00 ©2010 IEEE
7. Grewal, R.K., Pateriya, P.K.: A Rule-based Approach for Effective Resource Provisioning in Hybrid Cloud Environment. International Journal of Computer Science and Informatics 1(4) (2012)
8. Bobroff, N., Kochut, A., Beaty, K.: Dynamic placement of virtual machines for managing SLA violations'. In: Proc. International Symposium on Integrated Network Management 2007 (2007)
9. Verma, A., Dasgupta, G., Nayak, T.K., De, P., Kothari, R.: Server workload analysis for power minimization using consolidation. In: Proceedings of the 2009 Conference on USENIX Annual Technical Conference, San Diego, California, June 14-19, p. 28 (2009)

10. Clark, C., Fraser, K., Hand, S., Hansen, J.G., Jul, E., Limpach, C., Pratt, I., Warfield, A.: Live migration of virtual machines. In: NSDI 2005: Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation, pp. 273–286. USENIX Association, Berkeley (2005)
11. http://openstack-neat.org/
12. Calheiros, R.N., et al.: CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services, Technical Report, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia (March 13, 2009)
13. http://www.ics.uci.edu/~jutts/st108/Polynomial_and_Interaction_Regression_Models_in_R.doc

Multi-release Software: An Approach for Assessment of Reliability Metrics from Field Data

Varuvel Antony Gratus[1] and Xavier Pruno Pratibha[2]

[1] Aeronautical Development Agency, Bangalore - 560017, India
[2] Alcatel-Lucent India Ltd, Bangalore - 560045, India

Abstract. With the ever increasing demands in requirements and dynamic scenario, the complexity of the software logics and hence the programming increased multi fold. Owing to this, the likelihood of the fault introduction during the development stages also increased, in spite of adherence to the software coding & management standards, verification, validation and testing procedures adopted. Incorporation of software based processing is more pronounced due to the complexity, compactness and very short reaction time warranted. The effect of faulty software would be unimaginably severe, leading to catastrophic events, especially for the case of safety critical applications. Hence, it is essential to quantify the risk associated with software. For the purposes of risk quantification, the reliability metrics of the software, typically inherent failure rate, to be known a-priori. The estimation of field failure rates of software, which is of multi-release in nature, will solely, depends on the systematic collection, segregation and categorization of data. In this paper, an approach to carry out pre-statistical analyses of data is presented from the perspective of assessment of reliability metrics of software.

Keywords: Software reliability, Reliability Metrics, Failure Rate, Risk, Multi-Release Software.

1 Introduction

Growing demands of software systems in any field is justified for the vast user requirements and complex processing algorithms at the back end. As the systems are increasingly becoming complex and more capable, the usage of software processing is more profound in the realization process. Embedded systems are in the cutting-edge to cope with the present user requirements. Notwithstanding to that, embedded systems enable the user to collect, assess, interpret, manipulate, interact and implement data from/to more number of sources/destinations. Many industries are, hence, switching over to the computer based processing, detailing and production, wherein immediate real time processing and controlling of critical parameters are of essentiality. With the ever increasing demand of software based systems on the hand, it is essential to produce software which are risk-free and highly reliable. The OEM are at stake in demonstrating the

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 475–486, 2013.
© Springer International Publishing Switzerland 2013

reliability of the software by way of real time and/or simulated testing. The data collected during various phases of software testing and field trials are the baseline for establishing and demonstrating the software reliability. It is often assumed that, testing of software to happen, from the perspective of customer. This paper outlines the methodology of Reliability Centered Software Testing, Data Collection, Segregation and Categorization for the purposes of estimation of failure rate of software components, which is one of the critical software metrics. The outcome of this approach would essentially bring out data which are meaningful and s-significant.

1.1 Software Reliability

There are well established tools and techniques available to predict, estimate, assess and verify the hardware reliability. However, the same case is not applicable for software domain owing to many reasons. There are many models published in the literature for the prediction and estimation of software reliability. Every Software Reliability Growth Model [SRGM] estimates the parameters of reliability, closer to the reality, under varying assumptions and boundary conditions. The reliability of the software could be established by estimating the unknown parameters of the SRGM. This outcome of the software reliability model in terms of failure rate could be fed in, as input, to the Fault Tree Analysis [FTA] which forms the basis of any risk assessment.

1.2 Software Risk Assessment

Software Probabilistic Risk Assessment [SPRA] consists of a detailed analysis of the software realization. A thorough examination of functionalities with the requirements provides suitable starting point for the SPRA. Following are the steps involved in SPRA. The methodology for Software PRA is similar to the hardware PRA, except the prediction/estimation of occurrence probability of the initiating and basic events:

- Identify Hazards
- Explore the combination of possibilities which would eventually trigger the hazard identified
- Evaluate the risk using suitable techniques [Usually FTA]
- Verify the risk acceptance with the specified norms/safety standards or stated user requirements.
- List out the single point events
- Record the findings
- Enumerate the precautions, which could reduce the risk, if the evaluated value is higher. If not, suggest design improvements to mitigate the potential single point events.

2 Multi Release Software

Risk assessment of MRS is tedious due to the very nature of software development and testing. MRS is usually evolved from the predefined requirements and updated due to bug fixing. The code update was necessitated owing to the following reasons:

- Bug fixing
- Fine tuning of processing algorithms and limits (Improvement in performance, Fine tuning of levels)
- Change in requirements
- Change in functionality
- Change in external interfaces and algorithms

Accordingly, any software version may undergo n revisions, termed usually as *versions*. Refer [1], [2]. Testing would be carried out in all the versions of the released software versions. Module level testing is assumed to be *regression* type, whereas end-to-end integrated testing is assumed to be applicable irrespective of type and quantum of changes incorporated from version x to version $x+1$.

2.1 Reliability of Multi-release Software

As most of the present systems are being run in conjunction with software, Software Reliability is also an important factor affecting system reliability. But, software reliability differs from hardware reliability in that the former reflects the design perfection, rather than manufacturing perfection. The high complexity of software is the major contributing factor of Software Reliability problems. From the concept of MRS, it is to be noted that, every software version is released with some of the changes made from its previous versions. Developers tend to include more and more complexity into the software layer, with the rapid growth of system size and functionality requirement, by upgrading the software, as per the schema. Handling data from multi release software is cumbersome and a novel methodology is required to categorize and amalgamate the data together, wherever applicable. The realities of the field data usage on estimating the failure rate, are already discussed in [3] & [4]. Though the initial failure rate would start decreasing from the test start time of any release, it is assumed to increase during the next version, due to the new bugs introduced by the designers [imperfect debugging], by way of correcting the existing bugs and introduction of new features as given in Figure. 1.

2.2 Data Requirements for Estimation of SRM

Failure time data are essentially required for the purposes of estimation of Software Reliability Metrics [SRM]. Data would be in scattered form, particularly while dealing with multi-release software. The amalgamation of data becomes more and more complex, when the testing and field usage platforms of

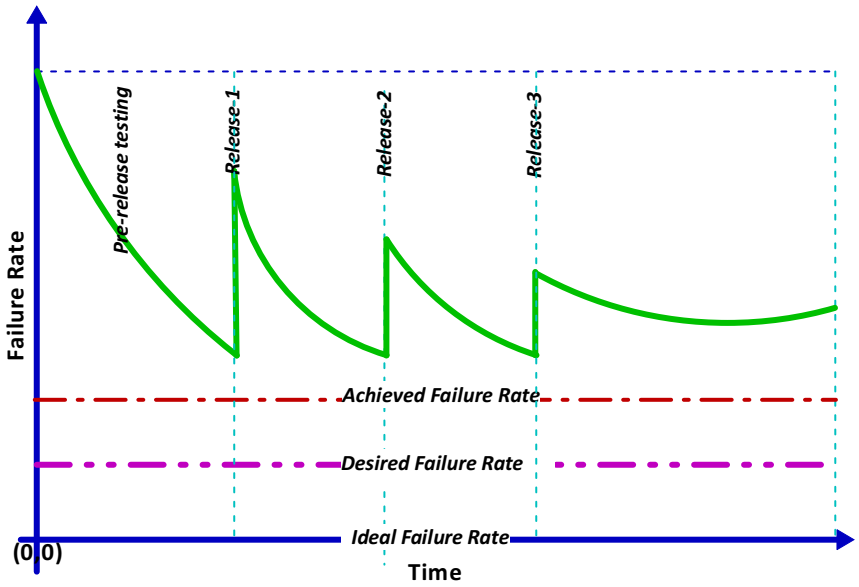


Fig. 1. Software failure rate vs time

the software, are many. This paper concentrates on the *time domain* approach of data. In totality, there are n software versions, and each software version is getting tested at m or lesser test platforms [TP]. The execution time of each software version is of prime importance. As an example for this case, two test platforms are assumed [Flight critical software x would get tested at *Integrated Testing facility*, prior to *usage on the aircraft*]. The software would get modified for the cases listed in Section. 2 on Page. 477, and the testing restarts, from a point of concern and interest. From the point of view of estimation of failure rate of MRS, which are getting tested at various TPs, following parameters are to be collected:

- Software Version
- Test Platform
- Total number of faults observed
- Hours of testing from previous fault [in the same TP]
- Total hours of testing

3 Data Analyses of MRS

Pre-statistical analyses of data is essential in filtering out *irrelevant* data from the set of data, which belongs to various software versions tested at various test platforms and the faults could be due to malfunctioning of hardware, software, user inputs, interfaces and human error. The following subsections describes in detail about the data segregation and categorization methodologies, application of an algorithm to group the *relevant* data.

3.1 Data Segregation Methodology

Data would provide meaningful information, if sampled and segregated properly considering the physical, functional and logical interrelations and importance of the same. In the case of software, voluminous data could get generated during the courses of design and testing. It is important to segregate the data which are *relevant* to the present problem at hand. Keeping the assessment of failure rate as the main focal point, the *relevant data* could be segregated from all available data. Data segregation of MRS would become more tedious, where the numbers of software versions are more and number of TPs, at which each software version is getting exercised, is also to be considered. With these, the total number of possible combinations would become a maximum of $m * n$. The segregation of relevant data involves:

- Unambiguous definition of error, fault and failure
- Validation of data
- Removal of faults and failures related/attributed to hardware design, hardware interfaces, workmanship and human error
- Identification of Software Version
- Relevancy check of the data
- Test Platform
- Hours of testing carried out on version n among m TPs.

The major outcome of the data segregation is the *Failure Time* of all the observed faults for all the software versions n with all the m TPs. The sample methodology of segregating *relevant* data is given in Figure. 2.

3.2 Data Categorization Methodology

In our present case, a software is assumed to be developed with modularity and every software version consists of k integrated and inter-related packages. There are two types of test platforms considered, for the sake of simplicity in solving the problem, viz., TP-1 & TP-2, which could further be extended to m TPs. The data segregated are to be analyzed for the applicability of software error. If so, the root cause of the fault to be identified to the malfunctioning/inconsistency/errors resident in specific package(s) of MRS and to be listed against the fault under the software version. Then the time to the occurrence of fault is calculated from the *Testing Log*. This time is non-cumulative and the time, the software was tested after its previous fault. i.e., t_{i-1} - t_i, where:

$$t_{i-1} : \text{Time to fault } i\text{-1}$$
$$t_i : \text{Time to fault } i$$

From this, the data segregated are categorized by package wise with the *Time tested to Fault* (This time is non cumulative, and indicates how many hours the software was tested for its occurrence from the previous fault). Consideration of modularity in development of MRS, the data categorization methodology should evolve a mechanism to identify and isolate the failures to module level/package level.

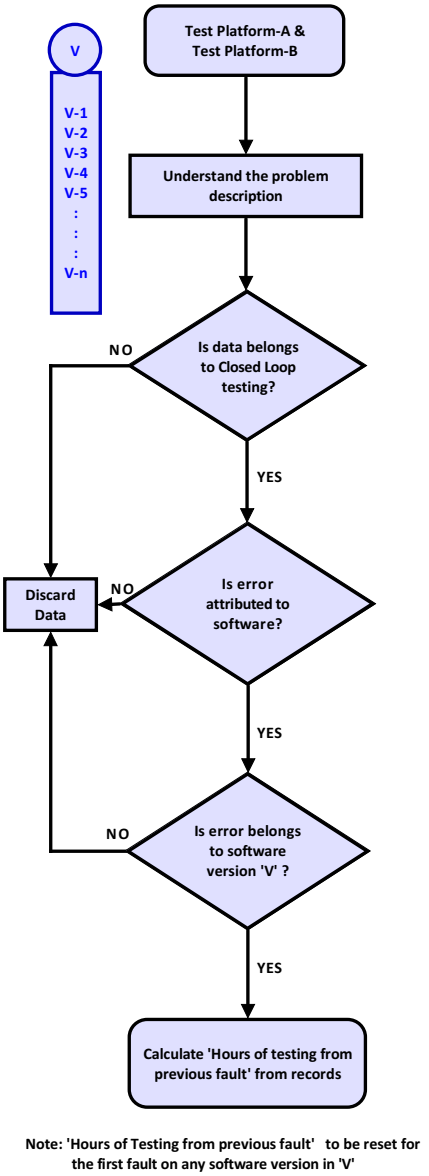


Fig. 2. Data Segregation

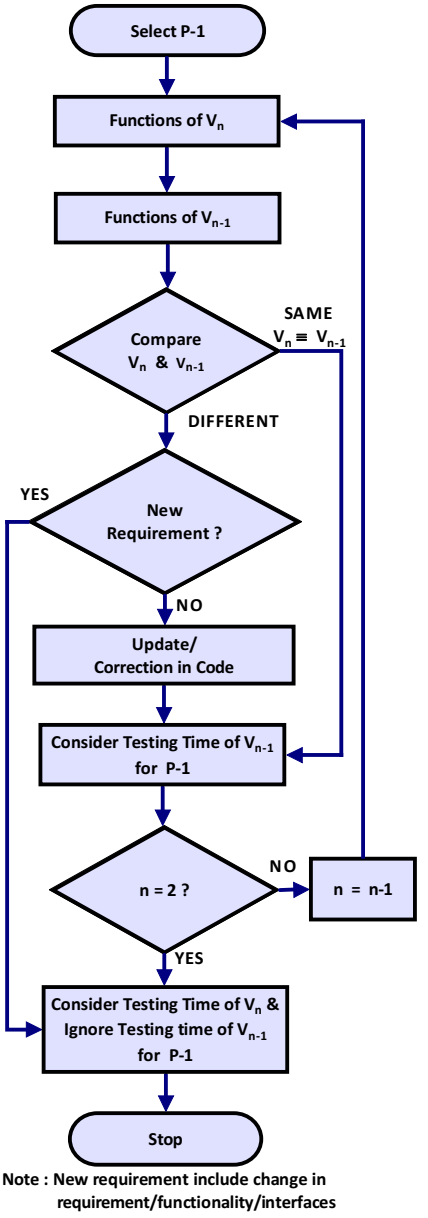


Fig. 3. Accounting testing time of multiple software versions for a package

Table 1. Packagewise Change log

Software Version	Package:1 Changed?	Package:2 Changed?	Package:3 Changed?	Package:4 Changed?	Package:5 Changed?	··· Changed?	Package:(m-1) Changed?	Package:m Changed?
Version: n	No	No	No	No	No	···	No	No
Version: n-1	No	Yes	No	No	No	···	No	No
Version: n-2	No		No	No	No	···	No	No
Version: n-3	No		No	No	No	···	Yes	No
Version: n-4	No		No	No	No	···		No
Version: n-5	No		No	Yes	No	···		No
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
Version: 5	Yes		No		No	···		No
Version: 4			No		No	···		Yes
Version: 3			Yes		No	···		
Version: 2					No	···		
Version: 1					No	···		

3.3 Data Amalgamation-Packagewise Change Log

With the basic assumptions of modularity and reusability, the packages are assumed to be modified in case of change in requirements/functionality/ interfaces. Due to the maintained and configured level of modularity, the technique presented in Table. 1 in Page. 481 is adopted. That is, the testing times of different software versions are combined in cases, where, additional requirements/interfaces/ functionalities are not added in the latest version, when compared to its previous software version. A flowchart presented in Figure 3 is useful in combining testing time of multiple software versions in respect of a specific package.

Based on the algorithm depicted in tabular form in Table. 1, the change history of every package has been analyzed starting from the latest version V:n to the first version V:1. In case, if there is any change in requirement/functionality/interfaces is/are introduced in any software version, then the testing time of that particular software version to the first version V:1 are excluded from the failure time and total time of operation of the particular package. That is, the package wise data are grouped, if the functionality remains same and unchanged across the software versions. Following guidelines are utilized while grouping the data belonging to various versions of software.

An application software version V:$n \neq$ V:n-1, when there:

- is a major change in requirement
- are new functionalities added to the software system (new features)
- is an addition of new hardware communication interfaces managed by software
- are addition of new packages introduced in the software
- is a complete rewriting of code in same language, or in other languages for some specific valid reasons is warranted (e.g., N-version programming)

Two MRS versions are treated as same, i.e., V:$n \equiv$ V:n-1, if and only if:

- Software corrections are carried out, to overcome the observed errors/faults/failures
- Fine tuning of computation/processing, based on feedback from testing
- Change in numerical bounds (Constants/Widening/Narrowing) for the purposes of observation or performance improvements
- Correction of erroneous logic coded.
- Bug-fixing
- Change in code, to adhere to coding standards

Using the methodology enumerated above, all those applicable data for a specific package could be categorized for further analysis. Refer Figure 3.

4 Failure Time Calculations

In the previous sections, it was dealt in detail about the methodologies in segregating, combining and amalgamating the data from different software versions for an identified software package. However, consideration of testing of software versions, at different test platforms, wherein there are no anomalies observed are also to be accounted for. Following are the cases applicable, in the present scenario:

- Error-free testing in TP-1 for version V:x
- Error-free testing in TP-2 for version V:x
- Error-free testing of versions V:$x+$ to V:-y in TP-1
- Error-free testing of versions V:$x+$ to V:-y in TP-2

The following subsections deal with all those possible combinations of accounting of error-free testing durations at different test platforms.

4.1 Accounting for Test Time without Failures

It is to be noted that, the package testing time involves multiple software versions and in multiple test platforms. This section dwells on the methods to account for, all applicable software version testing times, on all test platforms under the conditions that there were no failures observed during testing.

Following are the list of Testing Times which needs to be accounted for, for any package:

- Fault observed in TP-1, but not TP-2 for version x
- Fault observed in TP-2, but not in TP-1 for version x
- Fault not observed in versions $x+$ to -y, but observed in y, in TP-1
- Fault not observed in versions $x+$ to -y, but observed in y, in TP-2

Where,

x : Any software version with faults observed
y : Another software version with faults observed
$x+$: Software developed after x [succeeding version to x]
$-y$: Software developed before y [preceding version to y]

4.2 Effective Cumulative Time to Failure

Fault Observed in TP-1, But Not in TP-2 for Version x
To account for this error free testing time on TP-*2*, the error free testing time should be added to the test termination time at TP-1. The underlying reason for this is that, the trials on TP-2 are carried out after successful testing for the same version in TP-1. Hence, after testing in TP-2, if there are no errors observed on TP-2, which indicates that the software was tested for additional time in TP-2, during which no errors are found. This method is justified, due to the similarity of the operational profile in both TPs. Refer Figure. 4.

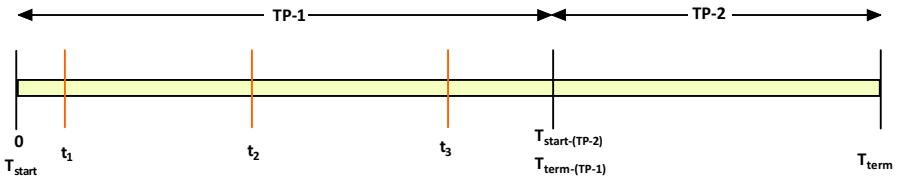


Fig. 4. Fault observed in TP-1, but not on TP-2

Fault Observed in TP-2, But Not in TP-1 for Version x
With the assumption that, TP-2 testing follows the TP-1 testing, if a software version tested for error free in TP-1, and then on TP-2, where faults have been found, it is required that, the Test Start Time could be modified accordingly to account for the error free testing time at TP-1. That is, the Test Start time on TP-2, should be the error free testing time in TP-1. Refer Figure. 5.

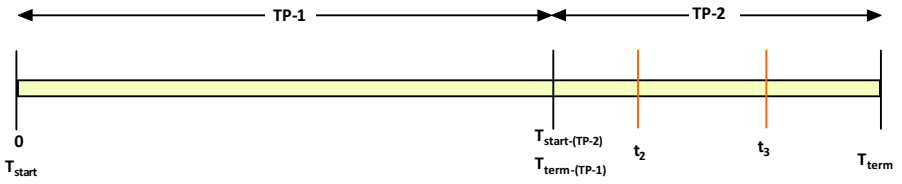


Fig. 5. Fault observed in TP-2, but not on TP-1

Fault Not Observed in $x+$ to -y, but Observed in x & y in TP-1
Special cases are also to be accounted for, as per the *Packagewise change log* in Table 1, that all the versions are to be considered for a specific package for the testing time, where there was no change in requirement in the software versions. Typically is the case, where there are faults observed in version x, and then with the continued test on higher versions up to -y, there was no errors found. But some other fault observed in version y. It could very well be appreciated that, the error free testing time between the versions with faults observed x and y,

could not be ignored. It could either be added to *Test Start Time* or to *Test Termination Time* without altering the physical interpretation of fault and the results contained therein and without affect the inter-fault time on any software versions. Following are the list of possibilities for any two subsequent software versions x & y. Let:

A : Test termination time of version x
B : Test termination time of version y
C : Error free testing time of version $x+$ to -y
D : Error free testing time of version x
E : Error free testing time of version y

Then, as per the above discussion, the test duration may get altered as: Test duration would be modified as:

$$T_{term(x_{EF})} = A + C + D \tag{1}$$
$$T_{start(y_{EF})} = T_{start(y)} + C \tag{2}$$
$$T_{term(y_{EF})} = B + C + E \tag{3}$$

Following points are to be considered while selecting the suitable options:

- Version $x+$ is updated when compared to x, with bug fixing (assumed to be perfect debugging), faults observed in x may not be observed in $x+$.
- Combining T_{EF} of higher versions with T_{start} of lower version, is, hence considered to be meaningless.
- In contrary, combining T_{EF} of lower version -y to T_{term} of higher version y will alter the physical interpretation of fault occurrence.
- Adding T_{EF} of higher versions to T_{term} of the previous version (with errors observed) conveys the correct interpretation of fault occurrence.
- As C is getting combined with T_{term} of x, it justifies that, faults found in x are corrected and hence those faults will not reoccur from $x+$ to -y. Physically meaning that, as if the test is continued with the lower version, and no anomalies found during the *extended testing time C*. [In other words, T_{EF} could be added to the T_{start} of higher version y]

Fault not Observed in $x+$ to -y, But Observed in x & y in TP-2
As per the discussions in para 4.2, combining T_{EF} of higher versions to T_{term} of lower version with faults, is considered applicable, irrespective of the test platform. The platform reference has no relevance in this case.

4.3 Time to Fault Calculation Methodology for MRS V-x

From the previous subsections, various possibilities have been discussed to account for the error-free testing time, while finding out the time to fault. Let

$T_{start_{(TP-1)}}$: Test starting time in TP-1 for MRS V-x
$T_{term_{(TP-1)}}$: Test termination time in TP-1 for MRS V-x
$T_{start_{(TP-2)}}$: Test starting time on TP-2 for MRS V-x
$T_{term_{(TP-2)}}$: Test termination time on TP-2 for MRS V-x

Fig. 6. Fault not observed in $x+$ to $-y$, but observed in x & y, in TP-1 or TP-2

Table 2. Failure Time Calculation

Failure in TP-1	Failure in TP-2	Time Correction
No	No	$T_{start_{(x)}} = T_{EF(-x)} + T_{start_{(TP-1)}}$
No	Yes	$T_{start_{(x)}} = T_{EF(TP-1)} + T_{start_{(TP-2)}}$
Yes	No	$T_{term_{(x)}} = T_{start_{(TP-1)}} + T_{EF(TP-2)}$
Yes	Yes	No correction required

Table 3. Packagewise CTTF - Template

Software Version x	Hours in testing	Data Source (TP)	Total hours of testing in TP	Version x : Error free testing in TP-1	Version x : Error free testing in TP-2	Version -x: Older Versions(s)-Error free testing time				
						In TP-1	Time	In TP-2	Time	Total time
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

From all the previous discussions and assumptions, the outcomes arrived are projected in Table 2, while accounting the *test time* and *error free test time* of *any software version* on *any platforms*. The methodology described in Table 2 & 3 for the calculation of CTTF of every fault could be programmed, to ease computations involving multi TPs.

4.4 Packagewise Cumulative Time to Failure

From the description presented in the previous paragraph, it is very much evident that, consideration of error free testing time is essential in order to obtain a better estimate of the failure rate. The rules described in the previous paragraph to be adopted, while evaluating the Cumulative Time to Fault for each and every package, for the observed fault. The template for the calculation of package wise cumulative time to failure [CTTF] is given in Table 3.

5 Conclusion

The approach presented in this paper was applied to the Flight Control Software which evolved from 1999 to 2012 with total number of versions 35 released during the period. Each software version consists of 25 packages [modules]. All the software versions have been tested in Integrated Test Facility prior to usage on aircraft. Of 35 versions, flight testing was carried out for 17 versions in one or more of the 7 flying aircraft. Integration testing which is in closed loop in respect of software packages, external interfaces and other associated inputs are resembling with the flight testing and hence, those data which are pertinent to IT and flight testing alone are considered. Failure rate have been estimated using the Lognormal Execution Time Software Reliability Model [5], based on goodness of fit test among the other models considered such as, Musa Execution Time Model and Binomial NHPP model. The software risk assessment in terms of estimation of probability of loss of control of aircraft has been computed using the FTA [Fault Tree Analysis] approach.

Acknowledgment. This work was sponsored by ADA-Bangalore and executed under the guidance of Reliability Engineering Centre, IIT-Kharagpur.

References

1. Hu, Q.P., Peng, R., Xie, M., Ng, S.H., Levitin, G.: Software Reliability Modelling and Optimization for Multi-release Software Development Processes. In: Proceedings of the IEEE IEEM (2011)
2. Kapur, P.K., Pham, H., Aggarwal, A.G., Kaur, G.: Two Dimensional Multi-Release Software Reliability Modeling and Optimal Release Planning. IEEE Transactions on Reliability 61(3) (2012)
3. Jeske, D.R., Zhang, X., Pham, L.: Accounting for Realities When Estimating the Field Failure Rate of Software. In: IEEE Transactions on Reliability (2001)
4. Jeske, D.R., Qureshi, M.A.: Estimating the Failure Rate of Evolving Software Systems. In: Proceedings of ISSRE 2000, 11th International Symposium on Software Reliability Engineering (2000)
5. Mullen, R.E.: The Lognormal Distribution of Software Failure Rates Application to Software Reliability Growth Modeling. In: Proceedings of the Ninth International Symposium on Software Reliability Engineering (1998)

On Emulating Real-World Distributed Intelligence Using Mobile Agent Based Localized Idiotypic Networks

Shashi Shekhar Jha, Kunal Shrivastava, and Shivashankar B. Nair

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam, India
{j.shashi,k.kunal,sbnair}@iitg.ernet.in

Abstract. Researchers have used Idiotypic Networks in a myriad of applications ranging from function optimization to pattern recognition, learning and even robotics and control. Most of the reported works that have used the Idiotypic network have been simulations wherein not all entities perform in a true distributed, parallel and asynchronous manner. The concentration of an antibody within the network is always assumed to be single valued, which is easily available as a global parameter in such simulated systems. This paper describes a novel architecture and dynamics to *emulate* an Idiotypic network wherein antibodies within a real physical network interact at antigen-affected nodes, sense their respective global populations stigmergically and form *Localized Idiotypic Networks* that eventually control their respective global populations across the network. *Typhon*, a mobile agent platform, running at the various nodes forming the physical network, was used for the emulation. While the mobile agents acted as antibody carriers and ensured their mobility, the nodes forming the physical network formed the antigenic sites. Results, portrayed herein, show the selective rise in global populations of the set of antibodies that are more effective in neutralizing a range of antigens across the network.

Keywords: Idiotypic networks, Emulation, Distributed Intelligence, Mobile agents, Typhon.

1 Introduction

The Idiotypic network model [1] which is inherently autonomous and has the ability of self-tuning, is a model which postulates that the antibodies interact with one-another even in the absence of an antigen. These interactions among the antibodies modulate the responses of the immune system as a whole. The formal mathematical model proposed by Farmer *et al.* [2] describes the concentration of an antibody to be affected by the amount of stimulations and suppressions it receives from other antibodies and antigens respectively together with the rate at which new ones are added and old ones removed. The Idiotypic network is a dynamic network which is regulated by the virtue of the concentrations of

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 487–498, 2013.
© Springer International Publishing Switzerland 2013

the various antibodies within the body. The concentration of an antibody refers metaphorically to its population in the system. In most AIS literature [3–6], these concentrations are always presumed to be single valued parameters. Further most of the implementations available for the Idiotypic network model are in the form of simulations [7] thus providing less room for its practical viability. To exploit the characteristics of the Idiotypic network model in real distributed systems, an architecture for the seamless interactions and operations of the concerned antibodies is crucial.

This paper presents a novel emulation architecture for realizing an Idiotypic network model over a real system of networked nodes which perform in a distributed and asynchronous manner. The novelty of our approach is that the intelligence is scattered in the environment (network of nodes) in the form of mobile agents [8] that act as antibodies, which selectively mitigate the problems arising at different nodes along with a competition among themselves to evolve the optimal solution. The succeeding sections provide a background on the related work, details of the proposed model for emulating an Idiotypic network over real systems followed by experimental results, discussions and conclusions.

2 Mobile Agents and Artificial Immune System

Mobile agents are autonomous chunks of software programs that can migrate within a network, carry payload, clone whenever required and terminate themselves if required [9]. These agents provide for a possible solution to emulate various population-based computational models in real systems. Using a mobile agent-based paradigm, Dasgupta *et al.* [10] describe a system for intrusion/anomaly detection and subsequent responses in networked computers. In his approach, the immunity-based agents roam around the nodes and routers monitoring the situation of the network. Inspired by the Clonal-Selection theory [11], the mobile immune agents used herein interact freely and dynamically with the environment and also with one another. Castro *et al.* [12], have proposed an artificial immune network model, for data clustering and filtering redundant data. They have used a Euclidean shape-space model in which the network units correspond to the antibodies. The input patterns were treated as the antigens to be recognized and clustered. This network model was successfully applied to several clustering problems, including non-linearly separable tasks. Godfrey *et al.* [13] describe an architecture of a multi-robot system that uses the AIS concepts and mobile agents to service robots. Based on pain, nodes that control the robot are triggered to indicate an antigenic attack. Mobile agents moving in a round-robin manner within the network carry the programs (antibodies) to decrease the pain levels of the robot.

3 Motivation for Idiotypic Network Emulation

Most of the systems that have used the Idiotypic network implement Farmer's [2] equation to deliver their models. These works are mostly simulations of the

Idiotypic network where the parameters involved are in some form accessible to all the entities in the network. The functioning of real systems, however remains grossly different and no real efforts seem to have been attempted to *emulate* Idiotypic networks on real networks. The biological Idiotypic network comprises several antibodies that stimulate or suppress one another and are generated based on their affinities with the concerned antigen. A stimulation causes the concerned antibody to increase its concentration i.e. its population increases since it has proved to be more effective in curtailing the antigenic attack. The opposite happens in case of a suppression whereby its concentration reduces. Successive suppressions may eventually lead to the removal of such antibodies. Thus, there need not actually be real physical link between all the individual members of the different antibody populations. At any moment of time during antigenic attacks, a distributed system could contain a repertoire of antibodies whose population sizes differ. If a specific antibody population seems more efficient in containing the antigenic attack its population (concentration) increases since the other less effective antibody populations stimulate it to grow. The more effective population may also suppress the growth of the other less effective ones.

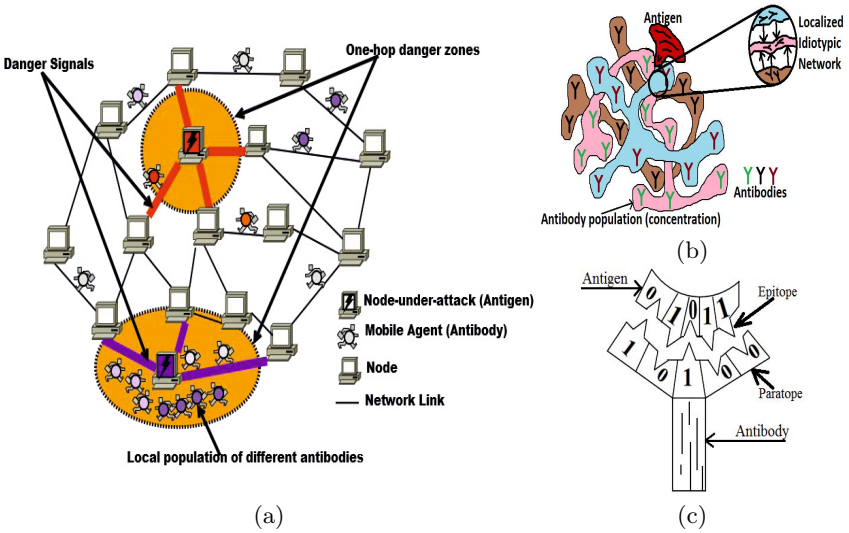


Fig. 1. (a) The proposed emulated Idiotypic model based architecture (b) An approximate visualization of the spatial distribution of antibody concentration and the localized Idiotypic network (c) Antigen and antibody in the proposed architecture

In the work reported herein, we have viewed the Idiotypic network as a network of populations (concentrations) of different antibodies. Each population communicates with the other using stimulations and suppressions, which cause dynamic changes in their respective populations thus contributing to a dynamic network. An approximate visualization of this network is shown in Figure 1(b).

As can be seen the populations of the different antibodies constitute a meta-level network but actual Idiotypic networks are formed at different spatial locations due to interactions of their sub-populations. What possibly is difficult to comprehend is the manner in which the idiotopes of all the antibodies of one population communicate and stimulate or suppress the others. Though in Figure 1(b) it seems that all the antibodies of one population stimulate all the others in another population via their idiotopes, this is not the way we envisage things in the work reported herein. We postulate that such stimulations happen only locally at the sites where an antigenic attack occurs. During an antigenic attack, the heterogeneous set of antibodies or sub-populations available in the locality of the attack, which are able cope up with the attack, compete with one another. The ones that are effective in containing the attack suppress those that are not, forming a *Localized Idiotypic Network* (LIN) in the locality of the attack. The sub-population of antibodies in this locality that performs better generate signals of suppression to reduce the number of the other sub-populations while the latter stimulate the former to increase the number of the more effective antibodies. The resultant effect is an increase in the number of the more effective antibodies in the locality of attack, thus containing the local antigen population quickly. Since antigens may attack in large numbers at different areas of a body, such small idiotypic interactions at these places add up to automatically increase the count of the more effective antibodies in the whole body. More effective antibodies are those solutions which have proved to be more effective against the problems or antigens. They also cause the numbers (populations) of other less effective ones to decrease, accordingly. All this happens in a stigmergic [14] and decentralized manner without all the antibodies of one kind interacting with all those of the others.

In the next section we describe the manner in which we portray a set of mobile agents acting as antibodies that move around within a real network of computers (nodes), finding and priming on antigenic attacks at the nodes and eventually increasing or decreasing their populations (concentrations) stigmergically, finally serving to *emulate* the Idiotypic network as described above.

4 The Emulated Idiotypic Network Model

The emulated Idiotypic network consists of a physical network comprising n nodes (computers) as shown in Figure 1(a). This set of networked nodes acts as the body of the system, *parts* (nodes) of which need to be *defended* or *serviced* by providing the best set of antibodies. A set of mobile agents move through this network of nodes and comprise (carry) the antibodies. Each node hosts a mobile agent platform to facilitate all mobile agent related functions including migration, cloning, antigen-antibody affinity measurements and generation of stimulations and suppressions. Antigenic attacks are initiated by presenting an antigen at the concerned nodes. Antigenic attacks can be viewed as a service required at a node while the antibodies that nullify these attacks could be seen as the relevant service providers. A node is thus the basic entity or the part

Table 1. Immune network metaphors in the Emulated Idiotypic network

Biological Immune network	**Entities within the Emulated Idiotypic Network**
Antigen	Service required at a node
Antibody	Mobile agent carrying services
Organs or parts of the body being defended	Nodes comprising the physical network
Antibody circulation	Mobile agent migration
Concentrations of various antibodies	Populations of the various mobile agents in the network
Stimulations/Suppressions	Increase/decrease in the sub-populations of concerned mobile agents within the node under attack
Increase/decrease in antibody count	Cloning of the stimulated mobile agents/ Termination of the suppressed ones, within the node-under-attack
Idiotypic Network formed by changes in stimulations and suppressions of antibodies in fluid (plasma) form	Dynamic changes in populations of each type of antibody due to stimulations and suppressions (received from others at nodes attacked by antigens), within the network

(organ) being defended or serviced within the system. Table 1 lists some of the mapping between the entities of the emulated Idiotypic network and their biological counterparts.

4.1 Antigen-Antibody Interactions at the Node-under-Attack

For the sake of explanation, we consider a random binary (m-bit) sequence to form an antigen which is presented at a node (node-under-attack or antigenic site). The binary sequence here is the representation of a problem at the node. The corresponding best antibody could be an m-bit complemented sequence capable of neutralizing the antigen. An antigen and its corresponding best antibody together with the epitopes and paratopes are shown in the Figure 1(c). In the proposed model, every mobile agent that acts as an antibody carries with it one neutralizing m-bit sequence. An affinity function ($\psi(A_g, A_i)$) defines the degree of interaction between the epitopes of an antigen (A_g) and the paratopes of the antibody (A_i).

$$\psi(A_g, A_i) = \frac{1}{(Epitopes\ of\ Ag)\ XNOR\ (Paratopes\ of\ A_i)} \tag{1}$$

The inverse of the XNOR distance between the bits corresponding to the epitopes and paratopes of an antigen and an antibody respectively describes the affinity of interaction between them. Hence, the best antibody would be the one, which has the complemented version of the antigenic epitope as its paratope.

4.2 Emulating Danger Signals at the Node-under-Attack

When a node is presented with an antigen (an m-bit string), it immediately radiates danger signals to its immediate neighbours which in turn diffuse the same onto their neighbours at a lesser intensity than that received. As shown in

Figure 1(a), these danger signals thus penetrate the immediate neighbourhood of the node-under-attack similar to the pheromone diffusion model proposed by Godfrey *et al.* [15]. In order to manage the diffusion within this sub-network each danger signal contains five parameters which include the identifier of the node-under-attack and the previous node, the epitopes of the antigen, a Diffused Signal Strength (DSS) whose intensity decreases as it diffuses to other nodes away from the node-under-attack and the life-time of the signal which also decreases similarly. The propagation of the danger signal continues till its strength dies down to zero at nodes in the neighbourhood of the node-under-attack, thus forming a danger zone around it as shown in Figure 1(a).

4.3 Antibody Migration and Generation

The mobile agents that represent the antibodies flowing in the network continuously migrate based on a combination of conscientious and danger signal oriented strategies, similar to that described in [15, 16]. The conscientious strategy ensures that the agents avoid recently visited nodes. However, when an agent detects a danger signal at a node, it ascertains whether it is a *candidate antibody* that can cater to the attack. This is done by calculating the affinity ψ between the neutralizing sequence carried by the mobile agent (antibody) and the epitope of the antigen within the danger signal. If this ψ is greater than χ, the affinity threshold (a non-zero positive constant), then the mobile agent assumes itself to be a candidate antibody and proceeds to tracking the increasing signal strength gradient towards the node-under-attack. This gradient aids the mobile agent to reach the node-under-attack via the shortest path [15]. This mechanism of attracting the relevant antibodies could lead to many candidate antibodies reaching the node-under-attack, some of which may be redundant. This redundancy is used to stigmergically sense the population of the candidate antibodies in the network. The numbers of each of the distinct candidate antibodies attracted to the node-under-attack is proportional to their respective global populations in the network.

If ψ is less than χ the mobile agents ignore the danger signals and continue to migrate to other nodes using the conscientious approach. It may be noted that only one out of the many antibodies that eventually reach the node-under-attack is chosen to neutralize the antigen. In addition, if the same type of antigen affects several nodes across the network simultaneously, it could be neutralized by different antibodies. It may also happen that the danger signals have died down due to the inherent lifetimes and no candidate antibodies have reached the node-under-attack. Under such a condition, the node itself starts generating antibodies proactively. In the present case, it generates random m-bit patterns and ascertains its ψ value. If the same is greater than χ then it uses this pattern (antibody) to neutralize the antigen. This new antibody is then encapsulated within a mobile agent and released into the network. This feature accounts for antibody generation within the network.

4.4 Stigmergy Based Antibody Population (Concentration) Control

In order to emulate Jerne's Idiotypic network [1], we have used a variant of Farmer's computational model [2] at each node-under-attack. The Farmer's equation in a general form can be written as:

Change in Concentration of an antibody = Antigenic Stimulation ($AgSt$) + Stimulations received from other antibodies (St) - Suppressions from the selected antibodies (Su) - Deletions due to disuse or lapse of Lifetime (Lt).

Whenever an antigenic attack occurs at a node (node-under-attack), the danger signal diffusions attract one or more antibodies, of the same or different types, to arrive at this node. Let ζ = {Type-1, Type-2,. . . . . .,Type-k} be the set of such distinct *candidate antibodies* that have arrived at the node-under-attack. Since multiple numbers of each of these types of antibodies could arrive at this node, each type of candidate antibody will have a population of its own within the node-under-attack. As can be seen in Figure 1(a) the node-under-attack at the bottom has three distinct types of candidate antibodies (shown in different shades) for the concerned antigen whose *local population* sizes are 2, 3 and 4 respectively. This is different from the *global population* sizes of the respective candidate antibodies which are not known to any single entity in the Idiotypic network emulation. Using the size of these local populations, the node-under-attack chooses that type of candidate antibody which has the highest local population as the best one for the neutralization of the antigen. We assume herein that in a distributed system, since more number of antibodies of this type have arrived at this node, the global population of this selected candidate antibody is high. Hence, it can be inferred that this type of antibody was possibly more effective in neutralizing attacks by this type of antigen at other nodes across the network. The node-under-attack thus senses the population size and decides the best candidate antibody by an indirect stigmergic manner. After the antigenic neutralization, the stimulations and suppressions received are used to increase or decrease the local population sizes of all the candidate antibodies at the node-under-attack based on an *activation factor* τ carried by each antibody within the network, details of which have been discussed later. These stimulations and suppressions create the *Localized Idiotypic Network* (LIN) among the local populations of the candidate antibodies at the node-under-attack some of which are shown and explained later in Figure 4. The LINs in turn alter the local population sizes increasing those that are stimulated and decreasing ones that are suppressed. These changes in the local populations (concentrations) at the node-under-attack contribute to the global ones and ensure that the more effective antibodies dominate the entire set of antibodies that flow in the network. It may be noted that in the emulated Idiotypic network, those mobile agents that carry such more effective antibodies, grow in number.

The equations that govern the dynamics of the formation of the LINs and the consequent changes in the local populations of candidate antibodies within the node-under-attack are given below. For all antibodies belonging to the local population, $a_i \in A_i$, the value of τ is given by -

$$\tau_{new}^{a_i} = \begin{cases} \tau_{old}^{a_i} + AgSt + St, \text{For selected candidate antibody population (Stimulation)} \\ \tau_{old}^{a_i} - Su, \text{For other candidate antibody populations (Suppression)} \end{cases} \tag{2}$$

where,

$$AgSt = \eta\psi(a_g, a_i) \tag{3}$$

$$St = \lambda_1 \frac{\sum_{x \in A_{Selected}} \tau^{a_x}}{\sum_{y \in A_{NotSelected}} \tau^{a_y}} \tag{4}$$

$$Su = \lambda_2\{\phi(A_{Selected}) - \phi(A_i)\} \tag{5}$$

$\tau \in [\tau_{min}, \tau_{max}]$
η = Antigen stimulation factor (non-zero positive value)
a_i = The i^{th} candidate antibody
a_g = Antigen at the node-under-attack
A_i = The antibodies forming the local population of the i^{th} candidate antibody that have arrived at the node-under-attack, $i \in \zeta$
$A_{Selected}$ = The local population of the selected type of candidate antibody used to neutralize the antigen
$A_{NotSelected}$ = The local population of those non-selected candidate antibodies
$\phi(A_i)$ = The population of set of antibodies A_i
λ_1 and λ_2 constitute the stimulation and suppression factors respectively which are positive non-zero values.

The ageing due to the Lifetime of the antibodies as mentioned in the Farmer's equation is handled separately. The change in population of the antibodies thus takes place by cloning or termination of antibodies (mobile agents) based on the following condition. For each candidate antibody a_i:

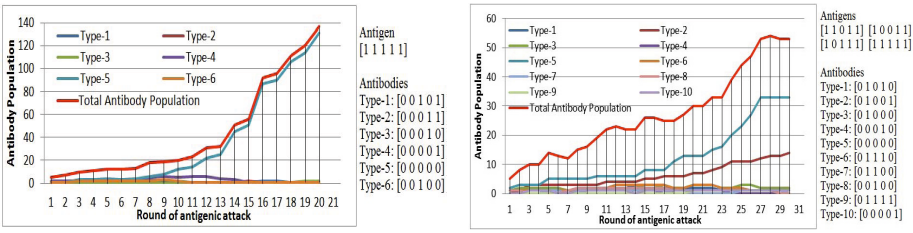
If $\{\tau^{a_i} \geq \tau_{max}\}$ ***then*** *clone* a_i, $\tau^{a_i} = 0$, $\tau_{clone}^{a_i} = 0$
Else If $\{\tau^{a_i} \leq \tau_{min}\}$ ***then*** *terminate* a_i

Since ageing is an integral part of an Idiotypic network, we have implemented this by conferring a fixed hop-count (H) to every mobile agent (antibody) in the network, which is reduced by unity at every hop. Once this hop-count becomes zero, the agents are terminated and hence removed from the system. In future, we intend to use concepts similar to that proposed in [17] to stigmergically control the agent population.

5 Experimentation and Results

The proposed model was emulated using *Typhon*, a mobile agent framework [18] on a 50-node network. Since we can instantiate multiple *Typhon* nodes on a single PC, the entire network was emulated using six PCs connected to each other via TCP/IP connections.

Initially since there were no antibodies in the network, the system used the method described in Section 4.3 to generate antibodies at various nodes-under-attack. At each of these nodes the concerned antibody was inserted as a payload



(a) (b)

Fig. 2. Antibody population in the system (a) when the same antigen was presented at five different nodes (b) when four different types of antigens were presented at five different nodes

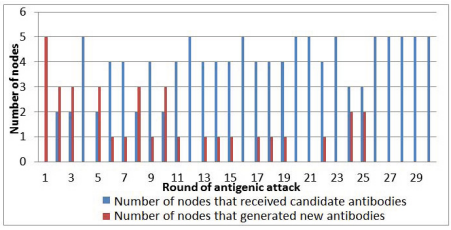


Fig. 3. Immunity of the system when four different types of antigens were presented at five different nodes

on to a mobile agent and empowered with $H = 100$, hop-lifetime which is an empirical estimate. Hence, each mobile agent carried a 5-bit string (a single antibody) as its payload. A 5-bit string was presented at a node to generate a node-under-attack. It must be noted that the antigens are the representations of problems occurring at a node and the antibodies are the corresponding solutions.

Each experiment performed consisted of multiple rounds of antigenic attacks. In each round, the system was made to be attacked by the same or different antigens at various nodes. The antibodies generated in each round were retained for use in the next. Experiments were performed by presenting antigens at various nodes, either simultaneously or consecutively. Results which highlight the effectiveness of the proposed architecture in a true distributed setting are presented.

Figures 2(a) and (b) show the variations of population (concentration) of each type of antibody along with the overall total population of antibodies in the system over several rounds. The antibodies generated and the antigen(s) presented are also shown within these graphs. When only one type of antigen was presented to five different nodes simultaneously for 20 rounds (20 attacks per node), six distinct antibodies (Type-1, Type-2, ...,Type-6) were generated across the network. It can be clearly seen that the population of the Type-5 antibody dominated the network while those of the others decreased drastically due to repeated suppressions at the various node-under-attack. Figure 2(b) shows a similar graph but the nature of antigen attack is different. Here, four different antigens were randomly presented at five different nodes for 30 rounds (30 antigenic attacks

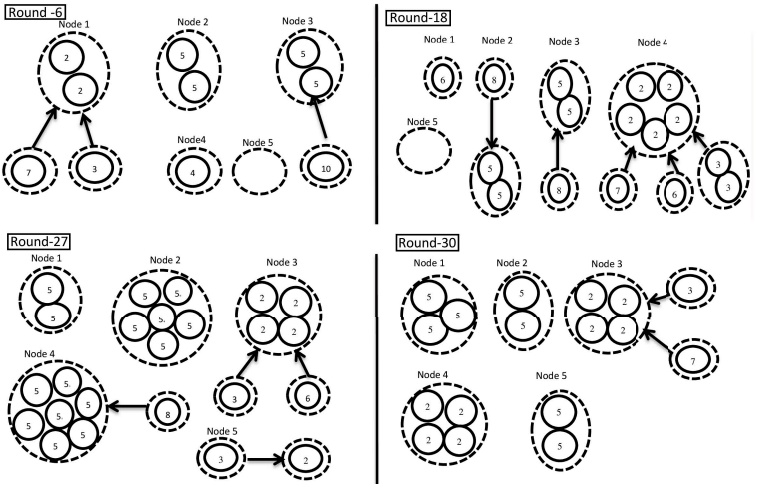


Fig. 4. Snapshots of the LINs formed during the rounds shown in Figure 2(b)

per node). It can be seen here that the Type-5 and Type-2 dominate the populations of antibodies. This possibly shows that these two types were capable of neutralizing all the four different antigens. This may also be verified from the bit sequences of the antigens and antibodies shown in the Figures 2(a) and (b). The ups and downs in the total antibody population in both the graphs shown in Figures 2(a) and (b) clearly indicate the regeneration and death of antibodies respectively.

Figure 3 depicts the manner in which the immunity of the network increases when four different antigens were made to attack five different nodes for 30 rounds as mentioned, in Figure 2(b). In the first round, since no antibodies populated the network, all 5 nodes-under-attack needed to generate antibodies locally as mentioned in Section 4.3. As the rounds increased, more antigenic attacks caused the generation of more effective antibodies that catered to some of the other nodes-under-attack. Eventually, beyond the 25^{th} round all nodes seemed to be catered to by the circulating antibodies and no new ones needed to be generated. For subsequent rounds possibly the populations of only two types of antibodies viz. Type-2 and Type-5 (see Figure 2(b)) were sufficient to neutralize attacks by the four distinct antigens. Figure 4 shows a few snapshots of the LINs formed during some of the rounds when four different antigens were presented at five different nodes in the network, as discussed earlier. Each circle with a solid boundary corresponds to a single antibody. The dotted boundary around the antibodies represent the local population of that type of antibody at the node-under-attack during the specified round. The number or the identifier within the antibody (solid circles) indicates the type number of the antibody. The empty dotted circles indicate that no antibody reached the node-under-attack. The arrow-heads indicate the direction of stimulations while their tails form the suppressions. It can be clearly seen that as the rounds progress the populations of antibodies of Type-2 and Type-5 grow and dominate the global

population in the entire network. At the 30^{th} round these two antibodies are the ones that have the highest sub populations at all the nodes-under-attack (viz. nodes 1 through 5) and are thus responsible for neutralizing the antigens at all the nodes.

6 Conclusions

In this paper, we discuss the manner in which the emulation of a real open-world model of an Idiotypic network on a physical network of computers, can be conceived. Results have shown how stigmergy based local interactions (stimulations and suppressions) at antigen-affected nodes can help generate Localized Idiotypic Networks of sub-populations of candidate antibodies, which in turn govern and control their respective global populations across the network thus validating the assumption made in Section 4.4. The emulation results also show how the populations of the more effective antibodies grow while the others die out and are thus removed from the network. The Idiotypic network can also generate new antibodies if required. The network also seems to be able to converge onto the more generic antibodies that are able to neutralize a set of varied antigens. Hence, given various solutions to solve similar problems arising in an distributed environment, the proposed architecture can evolve the optimal solution and purge the others.

We envisage that this emulation model, with customized modifications and improvements, will aid the realization of a plethora of real-world applications and aid AIS researchers to gain more insights into the actual distributed and parallel working of the Idiotypic network. We are currently working towards incorporating new features such as Clonal selection to evolve memory cells in lieu of the random method of generating antibodies, making some nodes act as lymph nodes and also vaccinating the network with antibodies which are known *a priori*, to eventually realize a networked *Artificial Being*.

Acknowledgements. The authors wish to thank the Department of Science and Technology, Government of India, for the funding provided under the FIST scheme to set up the Robotics Lab. (www.iitg.ernet.in/cse/robotics) at the Dept. of Computer Science and Engg., Indian Institute of Technology Guwahati, where the entire reported work was carried out and TCS for the financial support provided to one of the co-authors during the course of this research.

References

1. Jerne, N.K.: Towards the network theory of the immune system. Ann. Immunol.(Inst. Pasteur) 125, 373–389 (1974)
2. Farmer, J.D., Packard, N.H., Perelson, A.S.: The immune system, adaptation, and machine learning. Physica D: Nonlinear Phenomena 22(1), 187–204 (1986)
3. Watanabe, Y., Ishiguro, A., Uchikawa, Y.: Decentralized Behavior Arbitration Mechanism for Autonomous Mobile Robot Using Immune Network. In: Dasgupta, D. (ed.) Artificial Immune Systems and Their Applications, pp. 187–209. Springer, Heidelberg (1999)

4. Shimooka, T., Shimizu, K.: Idiotypic Network Model for Feature Extraction in Pattern Recognition – Effect of Diffusion of Antibody. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2774, pp. 511–518. Springer, Heidelberg (2003)
5. Hart, E., Ross, P.: Studies on the Implications of Shape-Space Models for Idiotypic Networks. In: Nicosia, G., Cutello, V., Bentley, P.J., Timmis, J. (eds.) ICARIS 2004. LNCS, vol. 3239, pp. 413–426. Springer, Heidelberg (2004)
6. Whitbrook, A.M., Aickelin, U., Garibaldi, J.M.: Two-timescale learning using idiotypic behaviour mediation for a navigating mobile robot. Applied Soft Computing 10(3), 876–887 (2010)
7. Greensmith, J., Whitbrook, A., Aickelin, U.: Artificial immune systems. In: Handbook of Metaheuristics, pp. 421–448. Springer (2010)
8. White, J.E.: Mobile agents. In: Software agents, pp. 437–472. MIT press (1997)
9. Outtagarts, A.: Mobile Agent-based Applications: a Survey. International Journal of Computer Science and Network Security 9, 331–339 (2009)
10. Dasgupta, D.: Immunity-based intrusion detection system: a general framework. In: Proc. of the 22nd NISSC, vol. 1, pp. 147–160 (1999)
11. De Castro, L.N., Von Zuben, F.J.: Learning and optimization using the clonal selection principle. IEEE Transactions on Evolutionary Computation 6(3), 239–251 (2002)
12. De Castro, L.N., Timmis, J.: Artificial immune systems: a novel approach to pattern recognition (2002)
13. Godfrey, W.W., Nair, S.B.: An Immune System Based Multi-robot Mobile Agent Network. In: Bentley, P.J., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS, vol. 5132, pp. 424–433. Springer, Heidelberg (2008)
14. Dorigo, M., Bonabeau, E., Theraulaz, G.: Ant algorithms and stigmergy. Future Generation Computer Systems 16(8), 851–871 (2000)
15. Godfrey, W.W., Nair, S.B.: A Pheromone Based Mobile Agent Migration Strategy for Servicing Networked Robots. In: Suzuki, J., Nakano, T. (eds.) BIONETICS 2010. LNICST, vol. 87, pp. 533–541. Springer, Heidelberg (2012)
16. Godfrey, W.W., Nair, S.B.: A bio–inspired technique for servicing networked robots. International Journal of Rapid Manufacturing 2(4), 258–279 (2011)
17. Godfrey, W.W., Nair, S.B.: A Mobile Agent Cloning Controller for Servicing Networked Robots. In: International Conference on Future Information Technology IPCSIT, vol. 13, pp. 81–85. IACSIT Press (2011)
18. Matani, J., Nair, S.B.: *Typhon* - A Mobile Agents Framework for Real World Emulation in Prolog. In: Sombattheera, C., Agarwal, A., Udgata, S.K., Lavangnananda, K. (eds.) MIWAI 2011. LNCS, vol. 7080, pp. 261–273. Springer, Heidelberg (2011)

Dependency-Based Query Scheduling in Distributed Data Warehouse Environment

Sakkarapani Krishnaveni and M. Hemalatha

Dept. Computer Science, Karpagam University, Coimbatore – 641 021
{sss.veni,hema.bioinf}@gmail.com

Abstract. A data warehouse is a representation of the elements and services of the warehouse, with particulars showing how the components will integrate with each other and how the usage of systems will grow over time. Finding the relevant information from a huge database is a hard task and consumes more time. This conflict is addressed using a query scheduling process in the data warehouse. In this paper, Dynamic Fault Tolerant Dependency Scheduling (DFTDS) algorithm has been proposed to schedule the queries based on their dependency and it automatically allocates the resources by checking the status of the virtual machine based on the acknowledgement of reply from client/user queries in distributed data warehouse systems. Experimental study of the proposed DFTDS algorithm shows a significant reduction in query processing time and memory utilization time compared with existing algorithms.

Keywords: Distributed Data Warehouse, WINE Algorithm, DFTDS Algorithm, Food Mart dataset, Consolidated Database System dataset.

1 Introduction

Distributed data warehouse looks like garbage collection gathered from multiple repositories for the purpose of analytical processing. Each data warehouse belongs to one or more organizations. The sharing of information can imply in the common format. When queries pose to the data warehouse environment, the query manager is directing queries to the appropriate tables. These queries are then mapped and sent to processors and execute the query reports by the use of task and resource scheduling algorithms. Scheduling queries are crucial task in the distributed data warehouse because of the amount of information and number of sites or resources or machines grows, adopting grid based task and resource scheduling algorithms are working well to elicitation of exact information. In that, existing algorithms are not producing the potential information from multiple repositories. To solve this crisis, we introduce Dynamic Fault Tolerant Dependency Scheduling (DFTDS) algorithm.

In Local Area Networks (LAN), distributed resources are configured with the physical machines (PM) or host above the PM, virtual machine (VM). In that, most aspect role in the machine is a VM. The virtual machine [7] typically limited in physical computing environment and make use of the components such as CPU usage, memory utilization, hard disk, network and hardware resources to be managed by a virtualization layer. These requests a transferred to the underlying physical

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 499–505, 2013.
© Springer International Publishing Switzerland 2013

hardware. In distributed data warehouse shared multiple repositories during the network that attempts to operate despite failure (Fault Tolerance). Fault Tolerance is used to prevent operation disruption due to hardware failures [2]. To avoid fault tolerant, our proposed approach enhanced with rollback recovery mechanism. Fault tolerance is achieved by frequently utilizing the state of a process during the failure-free execution and treats to communicate over a network via a collection of processes.

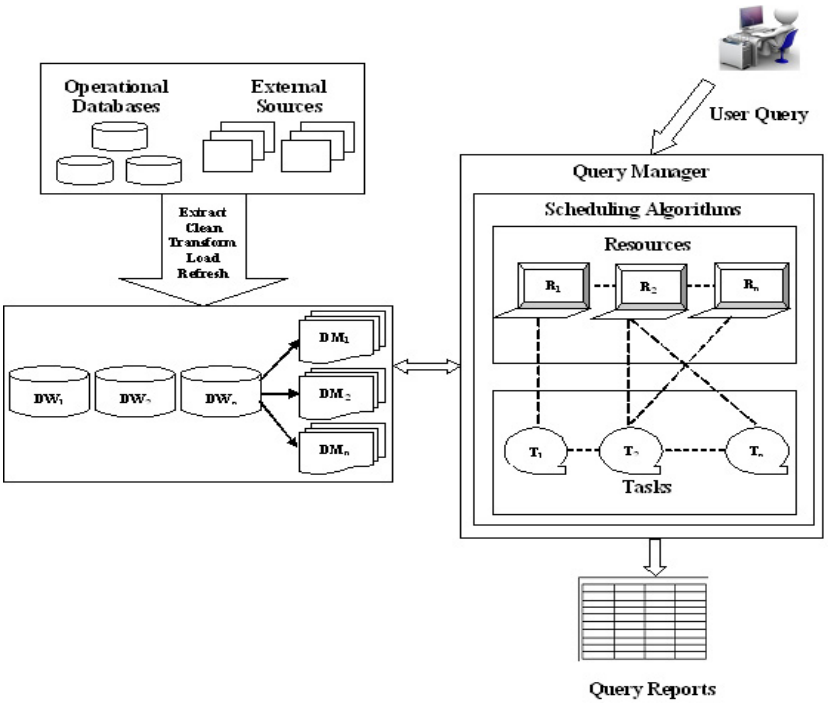


Fig. 1. Data Warehouse Architecture

In this paper, a new Dynamic Fault Tolerant Dependency Scheduling (DFTDS) algorithm has been proposed to solve the issues in the distributed data warehouse system. The proposed system integrates our previously proposed Dynamic Task Dependency Scheduling (DTDS) and Virtual Machine Fault Tolerant Resource Scheduling (VMFTRS) algorithms[3,4,5] for scheduling queries based on their dependency as well as recycling resources without human intervention in case of any fault occurred in the virtual machine. The fortifications are then evaluated through six parameters. A comparison of the proposed algorithm with Workload Balancing by Election (WINE) algorithm has been performed. WINE algorithm works well on clients' demands and provides better quality of service and data[6]. WINE is the two level scheduling algorithm. (a) Balance over the query and update queues. (b) Both queries and updates are prioritized based on quality of service and quality of data respectively. The Comparison results show that significant performance improvements can be gained through recovering and continuing after failure of virtual machines.

2 Proposed DFTDS Algorithm

The proposed DFTDS algorithm has been designed in such a way that initially user given queries are grouped based on their dependency as per DTDS algorithm and stored in query table. Physical machines (PM) and virtual machines (VM) with their bandwidth details are saved in the resource table. The grouped queries are allocated to resources based on VMFTRS algorithm followed by Virtual Computing Grid using Resource Pooling (VCGRP) technique[1]. This resource allocation detail is stored in allocation table and resource daemon reads this table to update an allocation table. Whenever queries are submitted to PM or VM an entry is made in submit table and query daemon reads the entries from this table to update the query table. All these four tables are maintained in resource distributing monitor. The work flow of the proposed DFTDS algorithm has been illustrated using a framework in fig. 2.

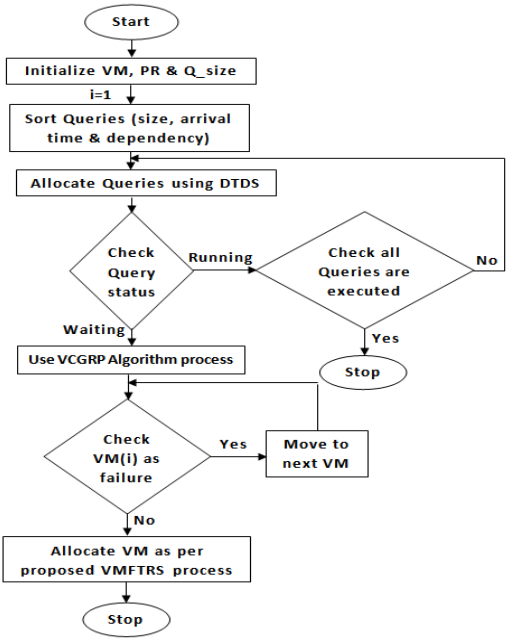


Fig. 2. Framework of Proposed DFTDS Algorithm

3 Dataset Description

Food Mart dataset (FM DS) and Consolidated Database System dataset (CDBS DS) has been used to evaluate the performance of the proposed DFTDS algorithm. FM DS contains twenty four relevant tables with 3,20,835 records. CDBS DS is a real time relational database[8] which contains forty six relevant tables with 87,43,045 records. An application id tuple is the primary key which links data elements for various tables and the files will be updated daily. In this work, both data sets are randomly distributed into different sites or machines.

4 Performance Metrics

To evaluate the performance of the proposed DFTDS algorithm various parameters have been used such as processing time, memory utilization, replication metric, error free execution, query cost and scalability. Formulation of the metrics are given below:

4.1 Processing Time (PT)

PT is the difference between the arrival time (t_{qi}) and the execution time (et_{qi}) of given queries. Where qi represents the (i) number of queries (q) given by the user.

Processing Time (PT_{qi}) = ($et_{qi} - t_{qi}$)

4.2 Scalability

The scalability is obtained by considering the processing time. Let $PT(Q_i(R_j))$ be the processing time of i queries in j resources. The average PT of the various sets of queries is evaluated by the same set of resources. Finally scalability is calculated as,

$$\text{Scalability} = \frac{PT(Q_i + + [R_j])}{PT(Q_i [R_j])}$$

4.3 Query Cost

The query cost is computed in terms of time. Let $PT_{q1,}$ $PT_{q2,}$….., PT_{qi} are the processing time of queries given by the user. Let dj be the data occurred in given queries.

$$\text{Query Cost (QC)} = \frac{1}{k}\sum_{k=1}^{n}\frac{PTqi}{dj}$$

where k represents the number of cycles same set of queries are executed. The unit of measurement is ms/data.

4.4 Duplication

Occasionally same queries are repeated in various sites or resources are defined as duplication. Let $(Qi \in Rj)$ as the i^{th} query in j^{th} resource.

$$\text{Duplication} = \frac{1}{k}[\sum_{k=1}^{i}(\sum_{j=1}^{n}[\sum_{j=1}^{m}(Qi \in Rj)])]$$

where k represents the number of cycles executed the same set of queries.

4.5 Success Rate

The success rate is calculated according to the total number of queries (Tq) from the user and the executed queries (Aq). The k iterations are made and evaluated the number of iterated errors (IEq) occurred. It is defined as the average rate of error occurred queries executed over different resources. In order to get the exact success rate, number of errored queries (Eq) calculated by Eq = (Tq – Aq). Then evaluate

iterated errored queries (IEq) by using $[\sum_{k=1}^{N}(Eq)k/N]$ formula. Finally success rate can calculate as,

$$\text{Success Rate} = \frac{[Tq - IEq]}{Tq} * 100$$

5 Experimental Analysis

To evaluate the proposed DFTDS and WINE scheduling algorithms using Java 1.7 for various given user queries. Both are online, non preemptive scheduling algorithms with different inclination. The performance of the proposed DFTDS and existing WINE methods have been used to evaluate by processing time, scalability, query cost, duplication and success rate performance metrics as well as food mart and CDBS datasets.

The processing time of DFTDS and WINE scheduling algorithms for two different datasets are evaluated by executing six different set of queries in the interval of 25 with the use of 10 resources. The aggregated results are shown in the following table 1 and fig. 3. For food mart dataset DFTDS algorithm takes 45, 48, 119, 131, 139 and 147 sec. for executing 25,50,75,100,125 and 150 queries respectively. For CDBS dataset DFTDS algorithm takes 129,195,310,407,528 and 666 sec. for executing 25,50,75,100,125 and 150 queries respectively. The results obtained show that when compared with a WINE algorithm our proposed DFTDS algorithm performs better.

Table 1. Processing Time (sec.) of Algorithms Vs Different Datasets

Number of Queries	FM DS		CDBS DS	
	WINE	Proposed DFTDS	WINE	Proposed DFTDS
25	60	45	136	129
50	68	48	205	195
75	174	119	314	310
100	181	131	424	407
125	193	139	540	528
150	224	147	708	666

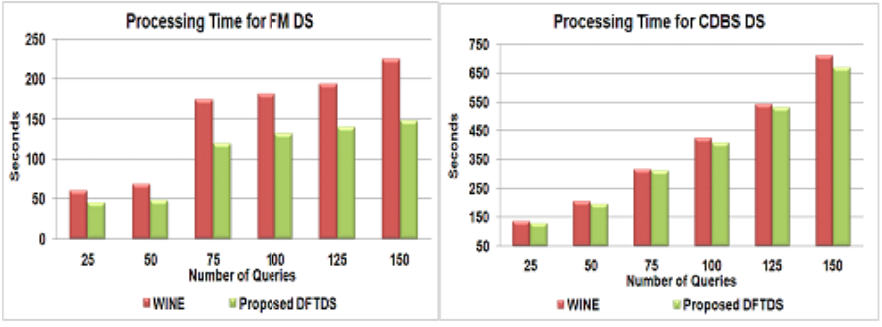


Fig. 3. Evaluation Plot of Processing Time

The scalability, query cost, duplication and success rate performance metrics for 25 cycles is measured. Each cycle evaluates the same set of 150 queries using 10 resources. The aggregated results are outlined in the following plot. The performance comparison of the proposed DFTDS algorithm in terms of scalability is 3.06 and 3.42 sec./query for FM DS and CDBS DS respectively. The query cost is 0.57 and 1.02 milliseconds/data for FM DS and CDBS DS respectively. The duplication is 7% and 4% for FM DS and CDBS DS respectively. The query cost is 0.57 and 1.02 milliseconds/data for FM DS and CDBS DS respectively. The success rate is 98% and 96% for FM DS and CDBS DS respectively. The final performance reports show that our proposed DFTDS algorithm performs better than an existing WINE algorithm.

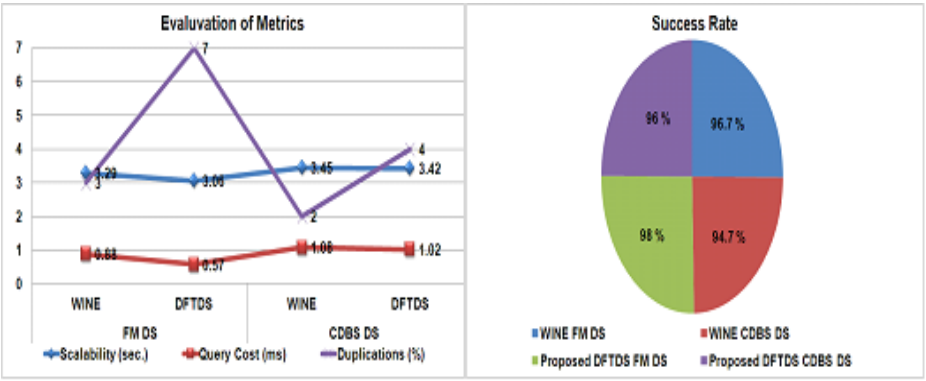


Fig. 4. Evaluation Plot for Algorithms

But sometime WINE scheduling algorithm could not produce the complete query result and could not send any error message when 100 queries were sent at a time. Hence, the proposed DFTDS algorithm has been executed on different query sets ranging from 25 to 150 with 10 resources. The lower processing time of DFTDS algorithm is obtained by automatic creation of new VM in case of any failure. Error rectification is performed within 5 seconds depending upon the size of information or data retrieved.

6 Conclusion

In this paper, the Dynamic Fault Tolerant Dependency Scheduling(DFTDS) algorithm has been proposed for the distributed data warehouse environment. DFTDS algorithm schedules queries by considering their dependency and resource status, then recycles the virtual machines to avoid the failures. Our proposed DFTDS is compared with an existing WINE scheduling algorithm under various parameters using different sizes of query sets and 10 resource groups. WINE is the existing data warehouse algorithm, based on queries and updates. For food mart dataset, the average performance evaluation in terms of scalability, query cost, duplication and success rate of our

proposed DFTDS algorithm is 7%, 35.2%, 57% and 1.3% outperforms than WINE. For CDBS dataset, the evaluation result in terms of scalability, query cost, duplication and success rate of the proposed DFTDS algorithm is 8.7%, 5.5%, 50% and 1.4% better than an existing WINE scheduling algorithm.

Acknowledgement. We thank Karpagam University for the Motivation and Encouragement to make this work as a successful one.

References

[1] Rajan, A., Rawat, A., Verma, R.K.: Virtual Computing Grid using Resource Pooling. In: IEEE-Int. Conf. Information Technology, pp. 59–64 (2008)
[2] Smith, J., Watson, P.: Fault-Tolerance in Distributed Query Processing, pp. 1–18 (2005)
[3] Krishnaveni, S., Hemalatha, M.: Query Processing in Distributed Data Warehouse using Proposed Dynamic Task Dependency Scheduling Algorithm. International Journal of Computer Applications 55(8), 17–22 (2012)
[4] Krishnaveni, S., Hemalatha, M.: Query Scheduling in Distributed Data Warehouse using DTDS and VMFTRS Algorithms. European Journal of Scientific Research 89(4), 612–625 (2012)
[5] Krishnaveni, S., Hemalatha, M.: Query Management in Data Warehouse using Virtual Machine Fault Tolerant Resource Scheduling Algorithm. International Journal of Theoretical and Applied Information Technology 47(3), 1331–1337 (2013)
[6] Thiele, M., Fischer, U., Lehner, W.: Partition-Based Workload Scheduling in Living Data Warehouse Environments. Information Systems 34, 382–399 (2009)
[7] Mohapatra, S., Smruti Rekha, K., Mohanty, S.: A Comparison of Four Popular Heuristics for Load Balancing of Virtual Machines in Cloud Computing. International Journal of Computer Applications 68(6), 33–38 (2013)
[8] The public files under CDBS are available at: `http://www.fcc.gov/mb/databases/cdbs/`

A Novel Bat Algorithm Based Re-tuning of PI Controller of Coal Gasifier for Optimum Response

Rangasamy Kotteeswaran[1] and Lingappan Sivakumar[2]

[1] Department of Instrumentation and Control Engineering,
St. Joseph's College of Engineering, Chennai India
kotteeswaranr@stjosephs.ac.in
[2] Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India
lingappansivakumar@gmail.com

Abstract. In this paper a novel BAT algorithm, a recently developed metahuristic algorithm is used to retune the parameters of pressure loop PI controller of coal gasifier, which is a highly nonlinear multivariable process having five controllable inputs and four outputs and strong interactions among the control loops. Functioning of coal gasifier involves many constraints to be satisfied on inputs and outputs. The existing controller along with its tuned parameters does not able to satisfy the constraints at 0% load for sinusoidal pressure disturbance and provides better response at 100% and 50% load conditions. The parameter of pressure loop PI controller is re-tuned using Lévy Flight (LF) guided BAT algorithm and performance tests which includes, pressure disturbance test, load change test and coal quality test are conducted. Test results shows that the re-tuned controller provides better response, meeting all the constraints at 0%, 50% and 100% load conditions.

Keywords: Bat Algorithm, ALSTOM benchmark challenge II, Coal gasifier, Integrated Gasification Combined Cycle, Metahuristic Algorithm, Lévy Flight.

1 Introduction

Integrated Gasification Combined Cycle (IGCC) produces clean power and energy with enhanced efficiency. It utilizes coal gas (also called syngas or producer gas) being produced from a processing unit called coal gasifier. Gasification is the process converting coal into coal gas under certain pressure and temperature. IGCC utilizes the purified form of coal gas to run the gas turbine to generate power and the exhaust gas exhaust gas from the gas turbine enters Heat Recovery Steam Generator (HRSG) to produce steam which in turn runs the steam turbine. And thus the efficiency of IGCC based power plant has increased efficiency compared to coal fired thermal power plants. Coal gasifier, an important and primary element in IGCC, is a highly non-linear, multivariable process, having five controllable inputs few non-control inputs and four outputs with a high degree of cross coupling between them. The process is a four-input, four output regulatory problem for the control design (keeping limestone at constant value). It exhibits a complex dynamic behaviour with mixed fast and slow dynamics and it is highly difficult to control and thus it is highly difficult to

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 506–517, 2013.
© Springer International Publishing Switzerland 2013

obtain the optimum response at the same time to meet the constraints for all operating points as given in the challenge problem [1]. The benchmark model of coal gasifier was developed by Alstom Power Technology, UK in two stages. Alstom benchmark challenge II includes a baseline PI controller which consists of three PI controllers and one feedback plus feedforward controller along with provision for conducting coal quality test. Until recently a group of researchers [2-10] have attempted to analyze and designed controllers and retuned the baseline controller to meet the performance objectives at all the load conditions. Apart from the conventional techniques, soft computing approaches such as MOGA [11] and NSGA II [12] are also used to design the controller. The recent developments in metahuristic algorithm inspire the authors to implement such algorithms to get optimum response.

2 Mathematical Representation of Coal Gasifier

Mathematically the transfer function model of the gasifier can be represented as;

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} & G_{13} & G_{14} & G_{15} \\ G_{21} & G_{22} & G_{23} & G_{24} & G_{25} \\ G_{31} & G_{32} & G_{34} & G_{34} & G_{35} \\ G_{41} & G_{42} & G_{43} & G_{44} & G_{45} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} G_{d1} \\ G_{d2} \\ G_{d3} \\ G_{d4} \end{bmatrix} \times d \tag{1}$$

Where,

G_{ij}=transfer function from i^{th} input to j^{th} output

y_1= fuel gas caloric value (J/kg); y_2=bed mass (kg);

y_3=fuel gas pressure (N/m^2); y_4=fuel gas temperature (K);

u_1 =char extraction flow (kg/s); u_2=air mass flow (kg/s);

u_3=coal flow (kg/s); u_4=steam mass flow (kg/s);

u_5=limestone mass flow (kg/s); d =sink pressure (N/m^2);

For a multivariable process decentralised control schemes are usually preferred. Equation 2 shows the structure of decentralised controller used in gasifier control [1]. It employs three PI controllers and one feedforward+feedback controller for coal flow rate.

$$G_c(s) = \begin{pmatrix} 0 & \left(K_p + \frac{1}{\tau_i s}\right) & 0 & 0 \\ K_f & 0 & K_p & 0 \\ 0 & 0 & 0 & \left(K_p + \frac{1}{\tau_i s}\right) \\ \left(K_p + \frac{1}{\tau_i s}\right) & 0 & 0 & 0 \end{pmatrix} \tag{2}$$

This baseline controller along with provided parameters satisfies the performance requirements at 50% and 100% operating points but fails to satisfy the constraints at 0% load for sinusoidal pressure disturbance(i.e. PGAS exceeds the limit of ±0.1bar). But this structure may give satisfactory response (meeting the performance requirements) if it is tuned optimally. The input and outputs should be maintained under certain limits for the proper operation of gasifier. The input actuator flow limits and rate of change of limit are associated with the physical properties of the actuator, should not exceed as shown in table 1.

Table 1. Input limits

Input variable	Max(kg s^{-1})	Min(kg s^{-1})	Rate(kg s^{-2})
Coal inlet flow (WCOL)	10	0	0.2
Air inlet flow (WAIR)	20	0	1.0
Steam inlet flow (WSTM)	6.0	0	1.0
Char extraction (WCHR)	3.5	0	0.2

Gasifier outputs should be regulated within the limits (table 2) for sink pressure (PSink) disturbance test, load change test and other tests. The desired objective is the outputs should be regulated as closely as possible to the demand.

Table 2. Output limits

Output variable	Desirable	Limits
Fuel Gas Calorific vale (CVGAS)		± 10KJ kg^{-1}
Bed mass (MASS)	Minimize	± 500 kg
Fuel Gas Pressure (PGAS)	fluctuations	± 0.1 bar
Fuel Gas Temperature (TGAS)		± 1 ºK

3 Bat Algorithm

Xin-She Yang[13], developed bat algorithm which is a population based metahuristic approach based on hunting behavior of bat. The following idealized rules are assumed for developing code for BAT algorithm [13] [14].

1. All the bats have the ability to identify and locate the prey by echolocation.
2. Bats flies with a frequency f_{min} from the current position x_i at a velocity v_i but with varying loudness and frequency.
3. The loudness varies from a minimum value(A_{min}) to maximum value(A_0)

Figure 1 shows the flow chart for Bat algorithm for the proposed tuning method. The wavelength (λ) and loudness (A_0) of bats varies to search for prey.

The frequency f_i and velocity v_i of i^{th} bat is updated by using the relation

$$f_i = f_{min} + (f_{max} - f_{min})\delta \tag{3}$$

$$v_i^t = v_i^{t-1} + \left(X_i^t - X_{gbest}^t\right) f_i \tag{4}$$

The new solutions can be found by

$$X_i^t = X_i^{t-1} + v_i^t \tag{5}$$

Where,

$\delta[0,1]$ =random vector from a uniform distribution.

f_{min} and f_{max} = function of domain size

Each bat is randomly assigned a frequency between f_{max} and f_{min}. This algorithm uses Lévy Flight [15] guided BAT. Each bat takes a random walk creating a new solution for itself based on the best current solution given by,

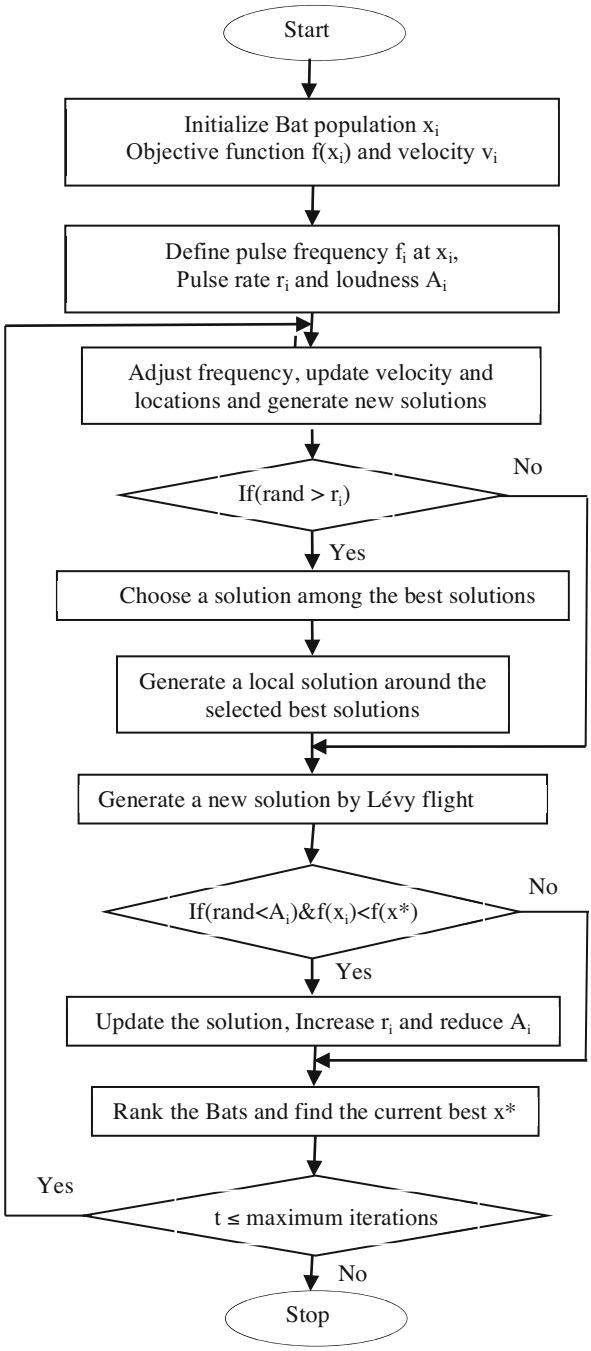


Fig. 1. Flow chart for BAT algorithm

$$X_{new} = X_{old} + A.|s|^{1/\beta} \tag{6}$$

Where,

$$A = \beta\Gamma(\beta)\sin\left(\frac{\beta\pi}{2}\right)\frac{1}{\pi}$$

β is the spatial exponent,
α is the temporal exponent, and
Γ(β) is a Gamma function

Loudness decreases as a bat move closer to its prey and pulse emission rate increases. Faster convergence is achieved with Lévy flight based randomization parameter.

$$A_i^{t+1} = \alpha A_i^t \tag{7}$$

$$r_i^{t+1} = r_i^0\left[1 - e^{-\gamma t}\right] \tag{8}$$

Where α and γ are constants.

4 Problem Formulation and Implementation

Figure 2 shows the implementation of BAT algorithm based optimization technique used for tuning the parameters of PI controller of coal gasifier. Maximum Absolute Error (MAE) for PGAS at 0% load is the objective function for BAT algorithm while the parameters of pressure loop PI controller (Pr_Kp and Pr_Ki) is the decision variables.

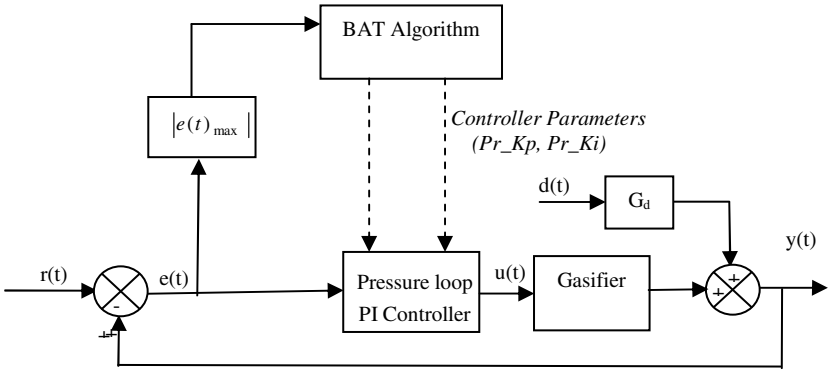


Fig. 2. Block diagram of Optimization scheme

Minimization of the desired performance specification is the objective function. The controller should respond quickly than the process and hence sampling time is selected as 1 seconds.

The procedure is as follows;

1) Initialize the variables and constants for Bat algorithm.
2) At 0% load and 0% coal quality apply a sinusoidal pressure disturbance (amplitude 0.2bar and frequency of 0.04Hz).

3) Run the simulation over 300seconds.
4) Calculate Maximum Absolute Error (MAE). This loop continues until the maximum iterations.
5) Best optimal controller parameters are obtained. These controller parameters of PI controller are the best tuned values for pressure loop PI controller.

Parameters of optimal PI controller and existing PI controller are listed in Table 3. These parameters are used to evaluate the performance of Optimal PI controller.

Table 3. Comparison of PI Controller parameters

Parameter	Dixon-PI[1]	BAT-PI
Pr_Kp	0.00020189	4.014882916673126e-4
Pr_Ki	2.64565668e-05	1.144954798223727e-7

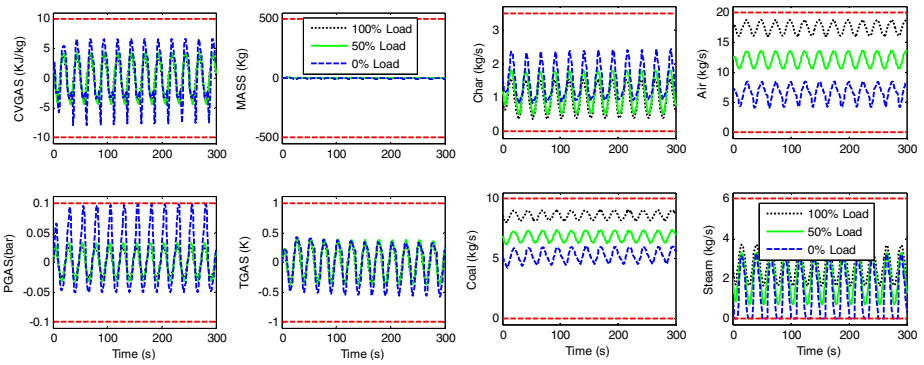
5 Performance Tests

Performance tests (pressure disturbance, load change and coal quality variation) are conducted to verify the applicability of the designed controller. The tuned controller parameters (Pr_Kp and Pr_Ki) for the pressure loop PI controller replace the existing parameters and the above mentioned performance tests are conducted. The gasifier input-outputs should satisfy the constraints mentioned in table 1 and table 2 for all performance tests.

5.1 Pressure Disturbance Tests

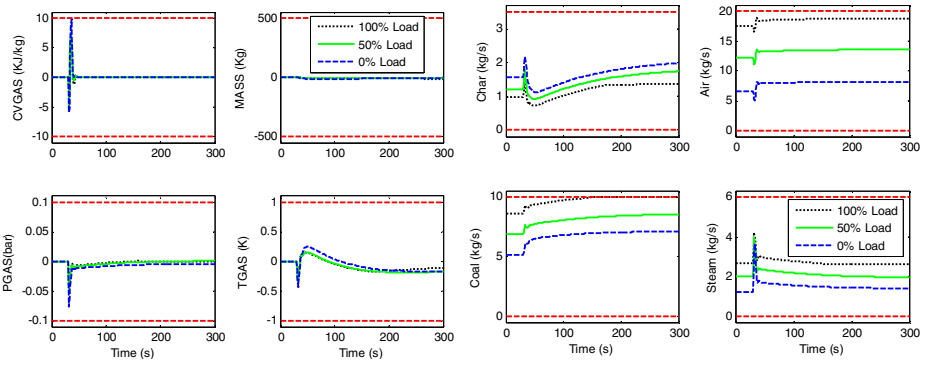
At 100% load, a sinusoidal pressure disturbance (amplitude 0.2 bar and frequency of 0.04Hz) is introduced and the response is monitored for 300 seconds. Integral of Absolute Error (IAE) and maximum Absolute Error (MAX) are calculated. This procedure is repeated for 50% and 0% load conditions. Figure 3a and figure 3b shows the deviation of the output from the setpoint and the manipulated inputs respectively for sinusoidal pressure disturbance at 0%, 50% and 100% load. On examining the inputs and outputs, it is noticed that the response meets all the performance specifications as mentioned in [1]. More particularly at 0% load condition PGAS is well below the limits but with the existing controller parameters [1], PGAS violates the constraints (exceeds 0.1bar limit). Similar experiment is conducted for step change in pressure disturbance and is shown in figure 4. It is clear that input-output response does not violate the constraints at all load conditions and for all disturbance tests.

The performance figures (IAE and MAX) for the above six pressure disturbance test are consolidated and are shown in table 4. Improved performance figure are realized for PGAS (indicated as bold figure) while a marginal increase in magnitude is obtained for the other outputs. This is due to the existence of strong interactions among the control loops. The response meets all the performance requirements at all the load conditions (0%, 50% and 100%) and for all the pressure disturbances (step and sinusoidal).



(a) Outputs and Limits (b) Inputs and Limits

Fig. 3. Response to sinusoidal disturbance at 0%, 50% and 100% load



(a) Outputs and Limits (b) Inputs and Limits

Fig. 4. Response to step disturbance at 0%, 50% and 100% load

Table 4. Summary of test output results

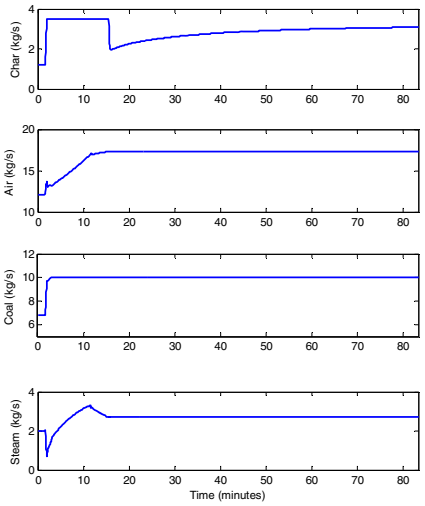
Test Description	Output	Maximum Absolute Error		IAE	
		BAT-PI	Dixon-PI[1]	BAT-PI	Dixon-PI[1]
100% Load, Step Disturbance	CVGAS(KJ/kg)	6.5038	4.8533	37.035	30.492
	MASS(kg)	6.9382	6.9383	768.72	795.29
	PGAS(bar)	0.0410	0.0499	0.6365	0.3883
	TGAS(K)	0.2941	0.2395	31.812	32.449
50% Load, Step Disturbance	CVGAS(KJ/kg)	7.3247	5.0310	38.920	32.224
	MASS(kg)	8.4548	8.4548	437.36	421.53
	PGAS(bar)	0.0494	0.0577	0.8552	0.4669
	TGAS(K)	0.3395	0.2660	37.278	38.433
0% Load, Step Disturbance	CVGAS(KJ/kg)	9.8772	5.8914	49.858	43.864
	MASS(kg)	11.053	11.053	580.91	667.90
	PGAS(bar)	0.0760	0.0772	1.9707	0.5949
	TGAS(K)	0.4315	0.3232	36.169	38.388

Table 4. (*Continued.*)

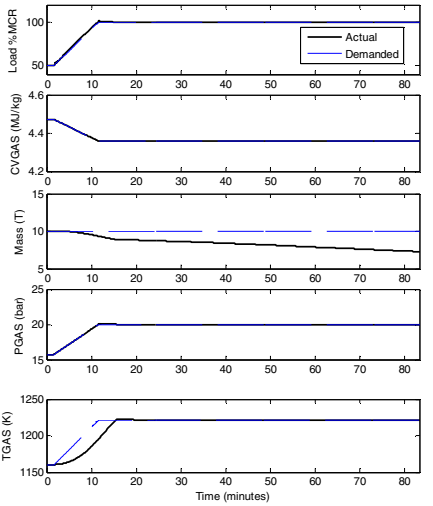
100% Load, Sinusoidal Disturbance	CVGAS(KJ/kg)	3.7838	4.1025	709.01	773.94
	MASS(kg)	10.799	10.858	2073.8	2076.5
	PGAS(bar)	0.0258	0.0496	4.8282	9.2825
	TGAS(K)	0.3566	0.3784	62.933	66.998
50% Load, Sinusoidal Disturbance	CVGAS(KJ/kg)	4.3730	4.7122	812.64	879.66
	MASS(kg)	12.783	12.852	2515.5	2522.2
	PGAS(bar)	0.0325	0.0623	6.0099	11.506
	TGAS(K)	0.3982	0.4226	70.358	74.717
0% Load, Sinusoidal Disturbance	CVGAS(KJ/kg)	7.9930	5.8585	1126.2	1039.5
	MASS(kg)	16.365	16.346	3018.1	3007.2
	PGAS(bar)	**0.0991**	**0.1196**	**12.702**	**19.145**
	TGAS(K)	0.5699	0.4791	85.674	79.541

5.2 Load Change Test

The stability of the system along with the controller is verified by conducting load change test. Here the gasifier with tuned PI controller is allowed to settle at 50% load and a ramp change (5% per minute) in load is applied. Response is recorded for 600 seconds (5% per minute). The actual load, CVGAS and PGAS track their demands quickly to setpoint while Bedmass takes more time to reach its steady state. The manipulated variables (coal flow and char flow) have reached their steady state immediately. Similar type of response is obtained when the process is started at 50% load and ramped it to 100% load(figure 5). This shows that the designed controller provides stable operation for ramp change in load.



(a) Input response to load change

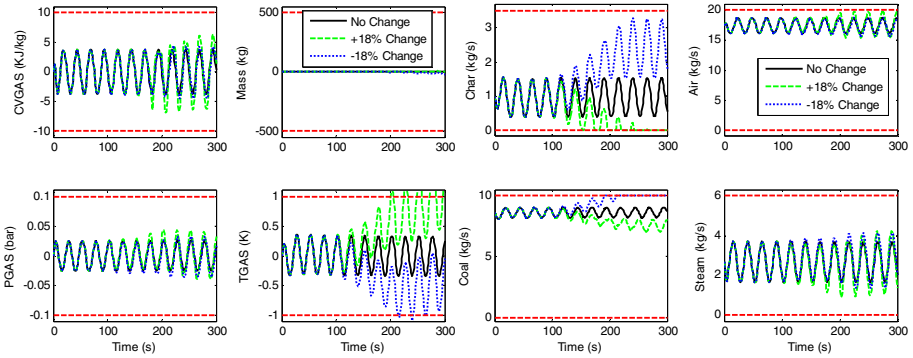


(b) Response to ramped increase in load

Fig. 5. Response to load increase from 50% to 100% load

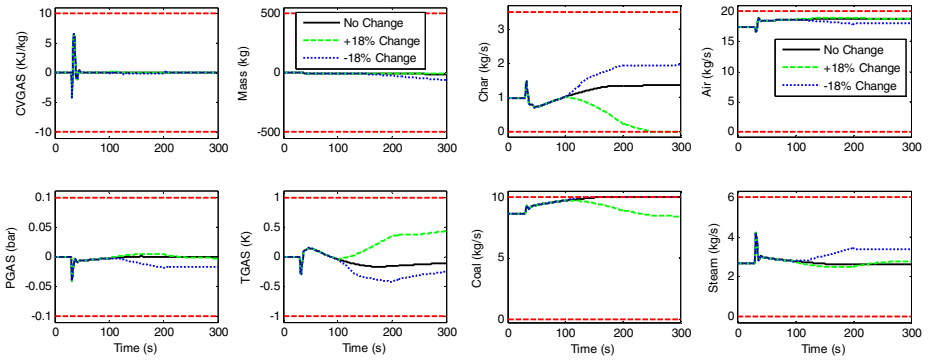
5.3 Coal Quality Test

Quality of coal greatly affects the production of syngas and hence the power production. Usually the coal quality is not constant over a period of time and may vary to a considerable amount. In this test, the quality of coal increased and decreased by 18% (the maximum possible change in coal quality), and the above pressure disturbance test are conducted to verify the robustness of the controller. Input-output responses (shown in figure 6-11) for sinusoidal and step change in pressure are verified for 300 seconds. The output figures indicates the deviation from the setpoint (i.e. error signal)



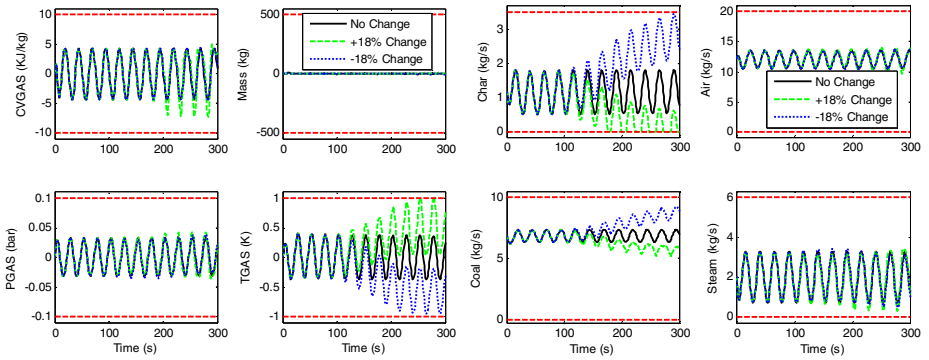
(a) Outputs and Limits (b) Inputs and Limits

Fig. 6. Response to change in Coal quality at 100 % Load for sinusoidal change in PSink



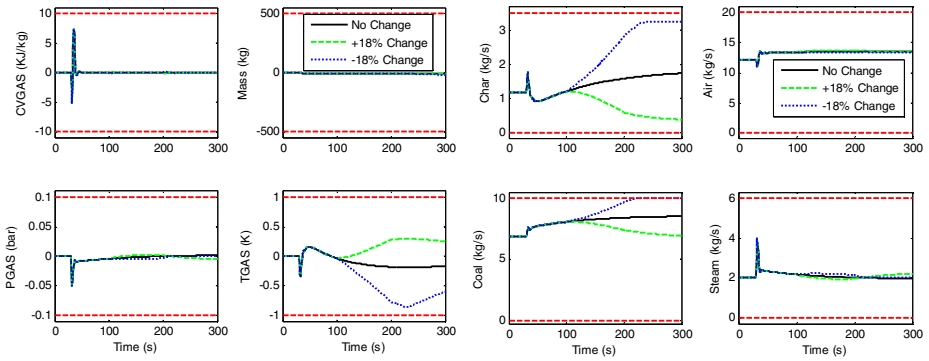
(a) Outputs and Limits (b) Inputs and Limits

Fig. 7. Response to change in Coal quality at 100 % Load for step change in PSink



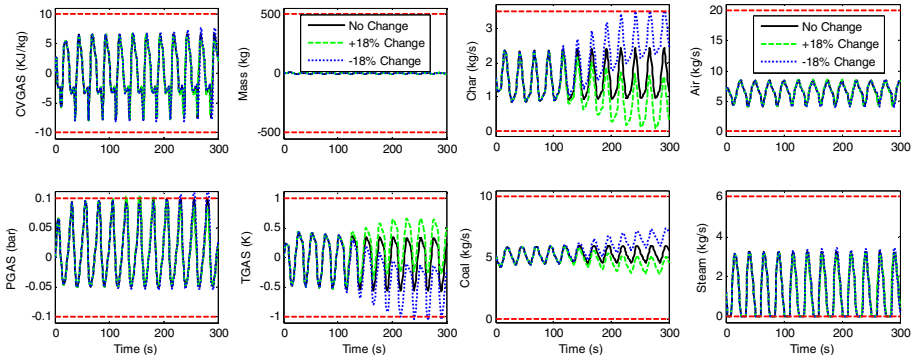
(a) Outputs and Limits (b) Inputs and Limits

Fig. 8. Response to change in Coal quality at 50% Load for sinusoidal change in PSink



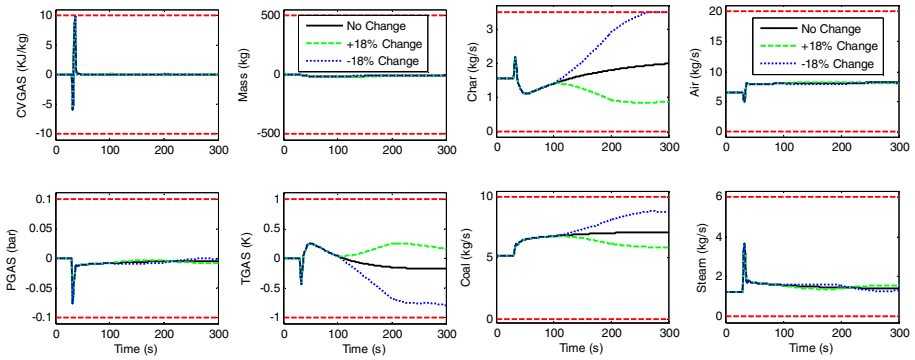
(a) Outputs and Limits (b) Inputs and Limit

Fig. 9. Response to change in Coal quality at 50 % Load for step change in PSink



(a) Outputs and Limits (b) Inputs and Limits

Fig. 10. Response to change in Coal quality at 0 % Load for sinusoidal change in PSink



(a) Outputs and Limits (b) Inputs and Limits

Fig. 11. Response to change in Coal quality at 0% Load for step change in PSink

The above figures show that all the outputs meet the performance criteria comfortably without violating the constraints. Table 5 shows the violation of the variables under positive (+18%) and negative change (-18%) in coal quality. Since input constraints are inbuilt in the actuator limits, output constraints are considered to be the actual violation. TGAS and PGAS violate the limits under change in coal in coal quality for sinusoidal pressure disturbance and no output variable is found for step pressure disturbance.

Table 5. Violation in output variables under coal quality change (±18%) (↑ - the variable reaches its upper limit, ↓ the variable reaches its lower limit)

Load	100%		50%		0%	
Disturbance type	Sine	Step	Sine	Step	Sine	Step
Coal quality increase (+18%)	Tgas↑	Within limits	Tgas↑	Within limits	Pgas↑	Within limits
Coal quality decrease (-18%)	Tgas↓	Within limits	Within limits	Within limits	Pgas ↑	Within limits

6 Conclusion

Metahuristic optimization algorithms are most widely used when the performance objective involves many constraints. In this work the baseline PI controller along with the controller constants violates the performance specification of the gasifier. The parameters of decentralised PI controller for pressure loop of Coal gasifier is retuned by using Lévy Flight (LF) guided BAT algorithm. The existing controller with tuned parameters does not satisfy the performance requirements at 0% load for sinusoidal pressure disturbance. Best optimum solution for the controller parameters are obtained for the given constraints and these parameters are used to get the desired response and also performance tests are conducted. Results show that this controller able to meet the performance requirements comfortably at 0%, 50% and 100% load conditions. And also the performance can be further improved by using Multi objective optimization algorithm.

Acknowledgement. The authors would like to thank Dr.Roger Dixon, Director of Systems Engineering Loughborough University, UK, Xin-She Yang, National Physical Laboratory, UK, for useful communication through email and Managements of St.Joseph's College of Engineering, Chennai and Sri Krishna College of Engineering & Technology, Coimbatore for their support.

References

[1] Dixon, R., Pike, A.W.: Alstom Benchmark Challenge II on Gasifier Control. IEE Proceedings - Control Theory and Applications 153(3), 254–261 (2006)
[2] Chin, C.S., Munro, N.: Control of the ALSTOM gasifier benchmark problem using H2 methodology Journal of Process Control 13(8), 759–768 (2003)
[3] Al Seyab, R.K., Cao, Y., Yang, S.: Predictive control for the ALSTOM gasifier problem. IEE Proceedings - Control Theory and Application 153(3), 293–301 (2006)
[4] Al Seyab, R.K., Cao, Y.: Nonlinear model predictive control for the ALSTOM gasifier. Journal of Process Control 16(8), 795–808 (2006)
[5] Nobakhti, A., Wang, H.: A simple self-adaptive Differential Evolution algorithm with application on the ALSTOM gasifier. Applied Soft Computing 8(1), 350–370 (2008)
[6] Agustriyanto, R., Zhang, J.: Control structure selection for the ALSTOM gasifier benchmark process using GRDG analysis. International Journal of Modelling, Identification and Control 6(2), 126–135 (2009)
[7] Tan, W., Lou, G., Liang, L.: Partially decentralized control for ALSTOM gasifier. ISA Transactions 50(3), 397–408 (2011)
[8] Huang, C., Li, D., Xue, Y.: Active disturbance rejection control for the ALSTOM gasifier benchmark problem. Control Engineering Practice 21(4), 556–564 (2013)
[9] Sivakumar, L., Anitha Mary, X.: A Reduced Order Transfer Function Models for Alstom Gasifier using Genetic Algorithm. Int. J. of Computer Applications 46(5), 31–38 (2012)
[10] Kotteeswaran, R., Sivakumar, L.: Lower Order Transfer Function Identification of Nonlinear MIMO System-Alstom Gasifier. International Journal of Engineering Research and Applications 2(4), 1220–1226 (2012)
[11] Griffin, I.A., Schroder, P., Chipperfield, A.J., Fleming, P.J.: Multi-objective optimization approach to the ALSTOM gasifier problem. Proc. of IMechE, Part I: Journal of Systems and Control Engineering 214(6), 453–469 (2000)
[12] Xue, Y., Li, D., Gao, F.: Multi-objective optimization and selection for the PI control of ALSTOM gasifier problem. Control Engineering Practice 18(1), 67–76 (2010)
[13] Yang, X.S.: Bat Algorithm for Multiobjective Optimization. Int. J. Bio-Inspired Computation 3(5), 267–274 (2011)
[14] Koffka, K., Sahai, A.: A Comparison of BA, GA, PSO, BP and LM for Training Feed forward Neural Networks in e-Learning Context. International Journal of Intelligent Systems and Applications 4(7), 23–29 (2012)
[15] Yang, X.S.: Firefly algorithm, Lévy flights and global optimization. In: Research and Development in Intelligent Systems XXVI, pp. 209–218. Springer, London (2010)

FI-FCM Algorithm for Business Intelligence

P. Prabhu[1] and N. Anbazhagan[2]

[1] Directorate of Distance Education,
Alagappa University, Karaikudi, Tamilnadu, India
[2] Department of Mathematics,
Alagappa University, Karaikudi, Tamilnadu, India
pprabhu70@gmail.com, anbazhagan_n@yahoo.co.in

Abstract. Business Intelligence combines the large data with analytical tools to present knowledge to the decision makers. It is used to understand the trends,future directions ,capabilities and technologies in the business. It has set of methods, process and technologies that transform raw data into meaningful information. Data Mining is one of Business Intelligence techniques that are used to obtain knowledge from data. The applications of business intelligence includes E-commerce recommender system, approval of bank loan, credit/debit card fraud detection etc., In this paper we have proposed FI-FCM algorithm for Business intelligence based on frequent itemsets and Fuzzy C Means clustering to extract the intelligence from the dataset in order to make the decision making process more efficient and to improve the business intelligence. E-commerce recommender system applications is selected to experiment this algorithm to help customers to find ,recommend products they wish to purchase by producing the list of recommended products.

Keywords: Business Intelligence, Frequent Itemset, k-means Clustering, Data Mining, Decision Making, Recommender system, E-commerce.

1 Introduction

In this real world, there exists a lot of information. It is necessary to maintain the information for decision making in business environment like approval of bank loan, e-commerce recommending systems, crime etc. The decision making is based on two kinds of large data such as OnLine Analytical Processing OLAP and OnLine Transaction Processing OLTP. The former contains historical data about the business from the beginning itself and the later contains only day-to-day transactions on business. Based on these vast data, decision making process can be carried out in order to discover intelligence.

Business Intelligence (BI) is a set of methods, process and technologies that transform raw data into meaningful and useful information. Data Mining is one the powerful new technologies in business intelligence with great potential that help the business environments to focus on only the essential information in their data warehouse. BI as the process of taking large amounts of data, analyzing that data, and presenting a high-level set of reports that condense the essence of

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 518–528, 2013.
© Springer International Publishing Switzerland 2013

that data into the basis of business actions, enabling management to make fundamental daily business decisions[11].In this paper, we have proposed FI-FCM algorithm based on frequent itemsets using FCM-clustering for business intelligence. E-commerce Recommender system applications is selected to experiment the algorithm.

1.1 Frequent Itemset Mining

Frequent itemset Mining is to find all the itemsets that satisfy the minimum support and confidence threshold. Support determines how often a rule is applicable to a given dataset, while confidence determines how frequently items in Y appear in transactions that contain X, These itemsets are called frequent itemsets used for finding intelligence by defining rules and clustering.

1.2 Cluster Analysis

Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. It plays an important role in scientific research and commercial application. There are various clustering algorithms have been proposed in the literature like k-means clustering, Partitioning Around Medoids (PAM) clustering ,Fuzzy C-means clustering, Hierarchical clustering, Density based methods etc. These clustering algorithms groups the data into classes or clusters so that object within a cluster exhibit same similarity.In this work Fuzzy C-means clustering is used for identifying similar objects for better decision making.

1.3 Recommender System

Recommender systems are one of the major important business intelligence systems. These systems are used by E-commerce websites to recommend the list of products to their customers based on the historical data. It improves the business in many ways like cross-sell by suggesting additional products,loyalty etc.In this paper,proposed algorithm is experimentally tested using recommender system.

2 Related Work

A new clustering technique which is extended from the technique of clustering based on frequent-itemsets[1]. Clustering based on frequent-itemsets has been used only in the domain of text documents and it does not consider frequency levels, which are the different levels of frequency of items in a data set. This approach considers frequency levels together with frequent-itemsets.

Given a large database of customer transactions, each transaction consists of items purchased by a customer in a visit[2]. They presented an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. They also presented results of applying the algorithm to sales data

obtained from a large retailing company, which shows the effectiveness of the algorithm.

The most popular frequent itemset mining called the FP-Growth algorithm was introduced[5]. The main aim of this algorithm was to remove the bottlenecks of the Apriori-Algorithm in generating and testing candidate set. The problem of Apriori algorithm was dealt with, by introducing a novel, compact data structure, called frequent pattern tree, or FP-tree then based on this structure an FP-tree-based pattern fragment growth method was developed. FP-growth uses a combination of the vertical and horizontal database layout to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item.

A general technique for applying Frequent and Space-Saving to transactional data streams for the case when the transactions considerably vary in their lengths[13]. Despite of its simplicity, author showed through extensive experiments that our approach is considerably more efficient and precise than the native application of frequent and Space-Saving.

The literature survey on cluster based collaborative filter and an approach to construct it is presented[11]. In modern E-Commerce it is not easy for customers to find the best suitable goods of their interest as more and more information is placed on line (like movies, audios, books, documents etc...). So in order to provide most suitable information of high value to customers of an e-commerce business system, a customized recommender system is required. Collaborative Filtering has become a popular technique for reducing this information overload. While traditional collaborative filtering systems have been a substantial success, there are several problems that researchers and commercial applications have identified: the early rater problem, the sparsity, problem, and the scalability problem.

Different item-based recommendation generation algorithms have been analysed [12]. They look into different techniques for computing item-item similarities (e.g., item-item correlation vs. cosine similarities between item vectors) and different techniques for obtaining recommendations from them (e.g., weighted sum vs. regression model). Finally, we experimentally evaluate our results and compare them to the basic k-nearest neighbour approach. Experiments suggest that item-based algorithms provide dramatically better performance than user-based algorithms, while at the same time providing better quality than the best available user-based algorithms.

3 Proposed Method

The aim of the proposed method FI-FCM is to identify and count frequent itemsets based on rules defined and cluster the objects to analyze the clustered objects in order to discover knowledge.This algorithm can be used to many

business to obtain knowledge from large datasets. The process involved in the proposed is summarized below:

1. Identify the dataset either synthetic or real world.
2. Choose the consideration columns/features
3. Define the rules
4. Count frequent itemsets satisfying rules
5. Cluster the objects
6. Validate the clusters
7. Discover intelligence

3.1 Identify the Dataset

Identify the dataset either synthetic or real-world like Credit/Debit card transaction, E-commerce, retail shop data, sanctioning of bank loan,crime etc.; Identified dataset is in the form;

$$D = \sum(A) = \{a_1, a_2, a_3, ..., a_n\}. \tag{1}$$

where, $\sum(A)$ is the collection of all attributes $a_1, a_2, a_3, ..., a_n$ are the attribute list that deals with the dataset D.

3.2 Choosing the Column/Features

Choose the considering column or features includes the elimination of the unwanted/irrelevant and redundant attributes / column / features / dimensions in the dataset. For feature selection heuristic methods such as Step-wise forward selection, Step-wise backward elimination, Combining forward selection and backward elimination Decision tree induction can be used. In Stepwise-forward selection the best of the original attributes is determined and added to reduced dataset[3]. At each step, it removes the worst attribute remaining in the set using Step-wise backward elimination method. In Combination forward selection and backward elimination method, the procedure selects the best attribute and removes the worst among the remaining attributes. Wavelet Transform and Principal Component Analysis can also be used for feature selection. In this work heuristic method is used for choosing the attributes / columns.

$$CC = \sum(A) - \sum(A'). \tag{2}$$

$$CC = \sum(A') = \{a_1, a_2, a_3, ..., a_n\} - \{a_{u1}, a_{u2}, a_{u3}, ..., a_{un}\}. \tag{3}$$

$$CC = D' = \sum(A') = \{a_1, a_2, a_3, ..., a_n\}. \tag{4}$$

Where, CC denotes the consideration column/features, which will be represented as (A'),(A) represents the set of all attributes in the chosen dataset, (Au) represents the set of all attributes to be eliminated to get the consideration column.

3.3 Defining Rules

The consideration columns consists of a list of attribute-value pair. From the consideration dataset D‘, the frequent itemset objects are identified based on some conditions that are defined in terms of rules. Sample rule can be defined as; $R = \{x | x \in D, x >= 20 \; and \; x < 60\}$

$\sum(R) = \sum(R)|X|\sigma((a_{ij}(A‘)/D))$

where, D represents the chosen dataset.

σ represents the selection process.

$|X|$ represents the join process.

$a_{ij}(A‘)$represents the consideration attribute of the new dataset A‘.

3.4 Count Frequent Itemsets Satisfying Rules Satisfying Rules

Count the number of occurrences of frequent itemsets that satisfies defined rules and threshold T(support and/or confidence etc.,).From the counted value, C_n, we can determine the number of frequent items that has been occurred in the dataset .This will generate the new dataset D“ with the above stated rules.

3.5 Cluster the Objects

The resultant dataset D“ is clustered based on the similarities using Fuzzy C Means clustering.Fuzzy C-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition[7].This is carried out by minimizing the objective function.

3.6 Validate the Clusters

Cluster validation is used to find the quality of clusters generated by the algorithm. It is performed by multiple simulations on a dataset varying the distance and clustering technique as well as the number of clusters k. There are three types of cluster validation to determine the optimum number of groups from a dataset, First way is to use external validation indexes for which a priori knowledge of dataset information is required, but it is hard to say if they can be used in real. Second way is to use internal validity indexes which do not require a priori information from dataset. Third way is relative index used to compare two different clusterings.

Since we do have priori knowledge about labels we can use internal indices like the Silhouette index, Davies-Bouldin index, Calinski-Harabasz index and Krzanowski-lai index to validate the clusters. The Silhouette index (SI)[4] validates the clustering performance based on the pair wise difference of between and within-cluster distances. In addition, the optimal cluster number is determined by maximizing the value of this index.For a given cluster, X_j

(j=1,..c), the silhouette technique assigns to the i_{th} sample of X_j quality measure, s(i)=(i=1,m),known as the silhouette width[6].

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{5}$$

where

a(i) is the average distance between the i^{th} sample and all samples included in the same cluster X_j;

b(i) is the minimum average distance between i^{th} sample and all samples clustered in other clusters.

The result of s(i) values are lies between -1 to 1. The value -1 indicates bad, 0 indicates indifferent and 1 indicates good. In our work Silhouette index (SI) is used to measure the quality of clusters.

The proposed FI-FCM algorithm for Business Intelligence is composed of the following steps;

Algorithm: FI-FCM

- Input: The number of clusters j and
  a dataset D with n objects.
- Output: A set of clusters C_j.

Begin

1. Identify the dataset,
   $D = \sum(A) = \{a_1, a_2, a_3, ..., a_n\}$ attributes
2. Outline the Consideration Column, CC from D;
   $CC = D' = \sum(A') = \{a'_1, a'_2, a'_3, ..., a'_n\}$ '
3. Repeat
4. Formulate the rules
   $\sum(R) = \sum(R)|X|\sigma((a_{ij}(A')/D))$, where i=1 to n rows, j=1 to m columns
5. Count the frequent itemsets
   C_n from $(d_{ij}(A')) > T$),where T specifies the threshold value.
6. Generate the Resultant Dataset, D" from C_n
7. Until no further partition is possible in CC.
8. Select initial centroids $C^{(0)}$
9. Initialize membership function
   $U = [U_{ij}]$ matrix,U^0
10. Calculate centre vectors
    $C^{(k)} = [c_j]$ with $U^{(k)}$
11. Update $U^{(k)}, U^{(k+1)}$ using degree membership.
12. If $max_{ij}||U^{(k+1)} - U^{(k)}|| < e$ then
    Stop
    else
    goto step 10;

End

The Figure 1 shows the process involved in the proposed Algorithm FI-FCM

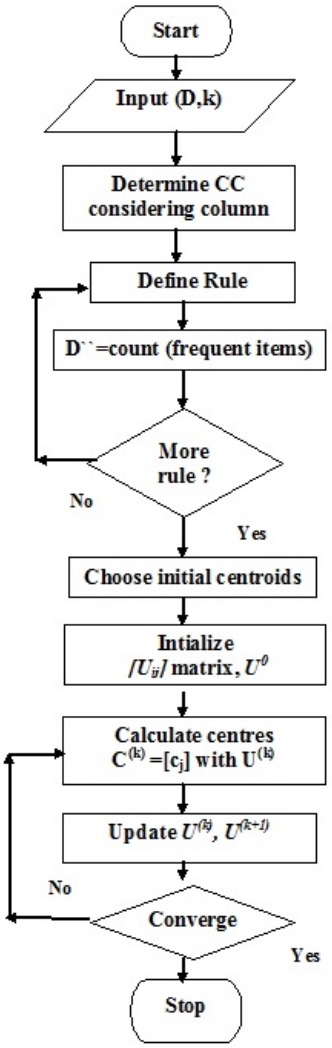


Fig. 1. Process involved in Proposed method

4 Experimental Setup

E-commerce recommender system application is considered to experiment FI-FCM Algorithm.This E-commerce purchase dataset contains 500 records of various products purchased by the customers visiting the E-commerce website. The Table 1 shows the description of historical synthetic E-commerce dataset. There

Table 1. Description of E-commerce Synthetic Dataset

Key Element	Description
Dataset Name	E-commerce
Original Attributes	Name,Gender,salary,date,quantity,item purchased,price and amount
Columns Considered	Age,quantity,price
Rules Defined	Age>= 20 *and age* <= 60, $Quantity > 2, Price > 1000, amount > 2000$

are some business E-commerce websites uses recommender systems like amazon.com,ebay.com,moviefinder.com etc.,

5 Result and Discussion

Many different algorithms like association rule based ,customer chosen rank based ,nearest neighbour based and most frequent item based have been applied for accurate and efficient business intelligence systems like recommender systems. In this paper frequent itemset count based FCM clustering algorithm FI-FCM have been proposed. This algorithm provides intelligence to the recommender systems by recommending most frequently purchased products to the customers .Based on the first rule defined, the new dataset D" can be formed with the details about the customer whose age lies between 20 and 60. .It can now be partitioned into 2 another sections based on count of frequent itemsets satisfying remaining rules defined. This dataset D" is clustered based on the similarities among the objects. These clustered objects gives intelligence to the customers and business.The number of objects allocated in each clusters using proposed method k=3 is tabulated in Table 2.

Table 2. Allocation of Objects in Clusters With k=3

k=3	Allocation of objects	Number of Objects
1	21,23,25,28,31-35,37,38,44-46,52,54,55,57-61	22
2	22,24,26,27,36,39,40,41,47,48,53,56	12
3	29,30,42,43,49,50,51	7
Total		41

There are 41 objects are identified for clustering, based on frequent items satisfying rules like age lies between 20 and 60. These objects are clustered using FCM clustering k=2 to 40 number of clusters.

The resulting k=3 cluster of objects calculated using FI-FCM algorithm is shown in figure 2. Here clusters are formed based on the similarities of objects. The validity of clustering is measured using SI silhouette Index. The figure 3 shows Silhouette plot for k=3 clusters.The SI values more close to 1 are clustered efficiently.

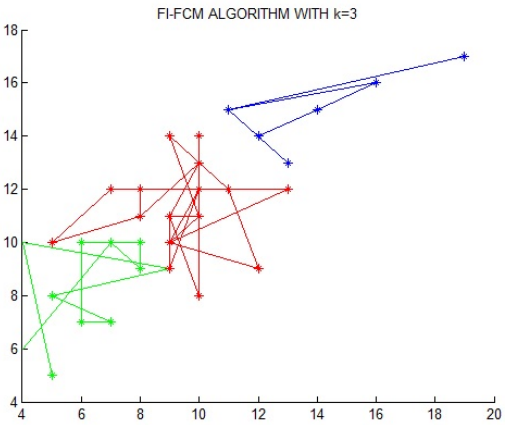


Fig. 2. Result of clustering using proposed algorithm with k =3

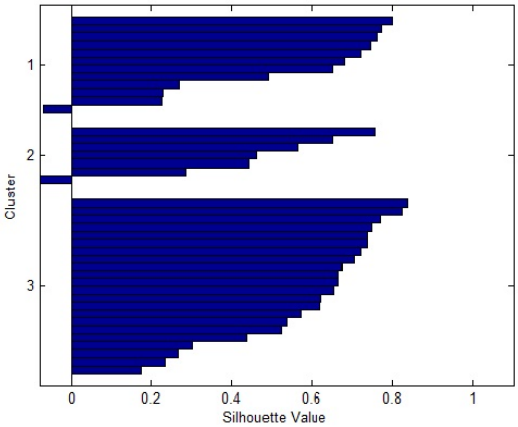


Fig. 3. Silhouette Plot for k=3

The Table 3 shows silhouette values calculated for k =2 to 5 number of clusters.

From the original dataset, identified products purchased by the customer belonging to this age group is shown in table 4:

In E-commerce recommender systems,when a customer enters into the site, the first step is to verify the customer details to identify, on which cluster the customer can be fall on. This can be done by analysing the features in the dataset. Here the feature, age of the customer is analysed. Based on the age, the customer can be easily classified and identified their position on the clusters. In our

Table 3. Silhouette Index Values for k=2 To 5

k	2	3	4	5
SI	0.41079	0.35334	0.36945	0.34871

Table 4. Identified Products Purchased by Customers

Age	Product Id	Frequently Purchased Product
20	IT100	Accessories
21	IT101	Computer
22	IT201	Jewels
23	IT303	Books
24	IT401	Sports items
.	.	.
60	IT110	Software

experiment, if the customer belongs to age 21 or 23 then they fall under cluster-1.Hence this customer can have more probability to purchase either computer or books. Thus, if the upcoming customer is identified his cluster as cluster-1 then proposed algorithm recommends product details of computers or books.if the customer belongs to age 22 or 24 then they fall under cluster-2.Hence this customer can have more probability to purchase jewels and sports items. Thus, if the upcoming customer is identified his cluster as cluster-1 then proposed algorithm recommends product details of computers and sports items. Hence this algorithm gives intelligence to the business and the customer by recommending the list of products instead of searching through various products.This will also reduce the searching time of customer and improves the business.

6 Conclusion and Future Work

This work shows that the proposed method FI-FCM based on frequent itemsets using FCM clustering is capable of identifying clusters in an effective manner to obtain the knowledge from data and to improve business intelligence for better decision making. Several experiments can be performed for performance benchmark that can be measured using Silhouette index. Other internal indices may also be used to test the performance of the clustering. More research work is needed to ensure its applicability in different domains like banking, risk analysis, tracking the performance of marketing campaigns etc.,.

References

1. Wimalasuriya, D.C., Ramachandran, S., Dou, D.: Clustering zebrafish genes based on frequent-itemsets and frequency levels. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 912–920. Springer, Heidelberg (2007)

2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases, IBM Research Center. In: Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA (May 1993)
3. Han, J., Kamber, M.: Data Mining concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers, San Fracisco (2006)
4. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20(1), 53–65 (1987)
5. Han, J., Pei, H., Yin, Y.: Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD 2000), Dallas, TX. ACM Press, New York (2000)
6. Rendn, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus External cluster validation indexes. International Journal of Computers and Communications 5(1), 27–34 (2011)
7. Al-Zoubi, M.B., Hudaib, A., Al-Shboul, B.: A fast fuzzy clustering algorithm. In: Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, pp. 28–32 (February 2007)
8. Stackowiak, R., Rayman, J., Greenwald: R Oracle Data Warehousing and Business Intelligence Solutions. Wiley Publishing, Inc., Indianapolis (2007)
9. Julashokria, M., Fathian, M., Gholamian, M.R., Mehrbod, A.: Improving electronic customers' profile in recommender systems using data mining techniques. Management Science Letters 1, 449–456 (2011)
10. Wach, E.P.: Automated Ontology Evolution for an E-Commerce Recommender,INFORMATIK 2011 - Informatik schafft Communities 41. Jahrestagung der Gesellschaft fr Informatik, Berlin (2011)
11. Venu Babu, R., Srinivas, K.: A New Approach for Cluster Based Collaborative Filters. International Journal of Engineering Science and Technology 2(11), 6585–6592 (2010)
12. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-Based Collaborative Filtering Recommendation, Algorithms, WWW 10, Hong Kong, May 1-5. ACM (2001) 1-58113-348-0/01/0005
13. Kutzkov, K.: Improved counter based algorithms for frequent pairs mining in transactional data streams. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part I. LNCS, vol. 7523, pp. 843–858. Springer, Heidelberg (2012)

An Algorithmic Formulation for Extracting Learning Concepts and Their Relatedness in eBook Texts

Rajesh Piryani[1], Ashraf Uddin[1], Madhavi Devaraj[2], and Vivek Kumar Singh[1]

[1] Department of Computer Science, South Asian University, New Delhi, India
[2] Department of Computer Science & Engineering, GBTU, Lucknow, India

Abstract. In this paper, we present an algorithmic formulation to automatically extract learning concepts and their relationships from eBook texts and to generate an RDF data that can be used for a number of purposes. Our algorithmic approach first extracts various parts of an eBook (such as chapters and sections) and then through a sentence-level parsing scheme identifies learning concepts described in the eBook text. We have programmed for the identification and extraction of relationships between different learning concepts occurring in a section. We have also been able to extract some general data about the eBooks such as author, price, and reviews (through eBook content mining and web crawling). The learning concepts, their relationships and other useful information extracted from the eBooks; is then programmatically transformed into a machine readable RDF data. The automated process of concept and relation extraction and their subsequent storage into RDF data, makes our effort important and useful for tasks like Information Extraction, Concept-based Search and Machine Reading.

Keywords: Information Extraction, Machine Reading, RDF Schema, Relation Extraction, Semantic Annotation.

1 Introduction

The fast and easily accessible Internet is driving new modes of information dissemination. Its universal connectivity and 'always on' access model is inspiring people to upload large and large amount of data and information in electronic form on the World Wide Web. It has affected all disciplines of life and teaching and learning is one such wonderful example. We now see various online course delivery platforms and a very large number of eBooks distributed through the Internet. The eBooks are rapidly becoming preferred mode of delivery of learning material, which can also be attributed to availability of cheap large screen eBook readers and devices. This new delivery model has reduced the cost of content dissemination as compared to the traditional book publishing model. However, the ease of creation, less costs involved in it and relatively not so stringent peer review system in eBook production model; makes it necessary to devise a method to identify an appropriate eBook for a topic and also to qualitatively evaluate the various eBooks available on a topic, something that have been attempted in [1], [2]. Further, a large number of eBooks pose the problem of information overload, when a reader seeks learning material on a particular topic/ theme.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 529–540, 2013.
© Springer International Publishing Switzerland 2013

It is with this motivation that we decided to devise an algorithmic formulation that can extract concepts described in an eBook so that we can associate semantic navigation tags to the eBook content. We have extracted concepts described in eBook chapters/ sections, along with the relations between learning concepts and other metadata from the eBooks. The information so extracted is transformed into a semantically navigable and machine readable RDF data. This allowed us to navigate the eBook collection and locate appropriate eBook and appropriate chapters in them for a particular concept of interest to the reader. All information extracted from eBook content and the one obtained from the Web, is the written programmatically in an RDF structure. The information thus stored allows us in identifying most suitable eBooks (and also relevant chapters/ sections) for a reader willing to learn some particular concept. The relevant information extracted from the eBook text is thus transformed into a machine readable form, which opens up many other useful application possibilities.

The rest of the paper is organized as follows. Section 2 describes the text parsing and concept extraction process. The concepts so extracted undergo a filtering process for identifying the CS domain only concepts. Section 3 describes the relation extractor we used and how it has been implemented. The relations identified are the ones which relate two learning concepts in CS domain. Section 4 illustrates the RDF structure, motivation for creating it and the process used for this. Section 5 describes the dataset used and the experimental results. This includes learning concepts extracted, relationship networks between learning concepts, RDF schema populated with data generated, and other information collected from external sources and made a part of RDF. The paper concludes in section 6 with a short summary and usefulness of the work reported here.

2 Concept Extraction

A learning concept usually refers to a distinctively identifiable piece of information/ concept to learn, understand and remember. It is a difficult task to parse the unstructured text of an eBook and identify which phrases correspond to a learning concept. We searched for any existing work that may help us in this process however; we could not find any directly applicable past work. We thought about new ways of doing so and found two approaches that may work. One of the simplest approaches could be to make a dictionary of all learning concepts in a domain and then use a string comparison to extract all phrases that match with any of the learning concepts in the dictionary. Though, simpler in logic, it is difficult to realize since constructing such an exhaustive dictionary is itself a very difficult task. Therefore, we tried to work on a generalized approach that is more robust and does not require naïve string matches. We found that notion of terminological noun phrases may work for our goal of extracting learning concepts. A past work [3] has shown that terminological noun phrases are good representative of concepts described in a text book. The only problem however is that it returns a large number of probable concept-like patterns, not all of which really correspond to a useful concept. We have therefore added a filtering process to this task to figure out relevant and useful concepts in CS domain from the large set of probable concepts.

2.1 Identifying Candidate Concept Phrases

First, we extracted all term phrases corresponding to the three linguistic patterns proposed in [4] and [5]. In order to identify and extract these patterns, it was necessary to parse the PDF eBook for its textual content and then process the eBook text sentence-by-sentence. The PDF eBook is transformed into usable text using iText API [6]. We also needed the chapter and section boundaries to be identified for each eBook so that it is possible to list the learning concepts belonging to a particular part of an eBook. Therefore, we extracted the bookmark information from the eBook and used the pagination information and text pattern matching to identify and extract various parts (preface, table of contents, chapters, sections etc.) for each eBook. Then we started the actual concept identification and extraction process for each chapter/ section.

Given a chapter, we process the text of the chapter sentence-by-sentence to identify the concepts. For each sentence, we had to run a POS tagger[1] to identify patterns with a particular linguistic nature and hence assumed to be part of a terminological noun phrase. The tagger assigns a unique part of speech to each word in a sentence we look for information represented by a set of words following a specific part of speech pattern. We parse each sentence and look for the three word patterns:

```
P1 = C*N
P2 = (C*NP)?(C*N)
P3 = A*N+
```

where **N** refers to a noun, **P** a preposition, **A** an adjective, and **C = A** or **N**. The pattern **P1** represents a sequence of zero or more adjectives or nouns which ends with a noun and **P2** is a relaxation of P1 that allows two such patterns separated by a preposition. The '*****' symbol represents any number of occurrences of **C** and '**?**' represents that two patterns may be joined through a preposition. Examples of the pattern **P1** include 'probability density function', 'fiscal policy', and 'thermal energy'. Examples of the **P2** pattern include 'radiation of energy' and 'baud-rate of modem'. The pattern **P3** corresponds to a sequence of zero or more adjectives, followed by one or more nouns. In **P3**, an adjective occurring between two nouns is not allowed that means it is restricted version of **P1**. Candidate concepts always comprise of maximal pattern matches i.e., if the actual word pattern is "probability density function", the likelihood of reading it as concept "probability density" is negligible. The target is to identify more specific concepts than general concepts. We have used all the three patterns as representative of concepts. This resulted in a large number of concepts being extracted. Since the goal was to extract relevant and useful concepts in CS domain, the large list of probable learning concepts was subjected to a filtering/ pruning process that is able identify useful concepts.

2.2 Filtering Valid Concepts from Candidate Patterns

Since the target is to identify and extract only CS domain learning concepts, the candidate set of concept patterns is subjected to a filtering/ pruning process. This involved matching the entries in the candidate concept list with an augmented ACM Computing

[1] http://nlp.stanford.edu/software/tagger.shtml

Curricular Framework (CCF) document that we prepared for this purpose. The CCF document defines the CS body of knowledge and classifies it into 14 different categories. This information was used as a reference and CCF constitutes the base reference document for concept filtering. To make the filtering process more accurate, three standard reference documents, namely ACM Computing Curricular framework[2], IEEE Computer Society Taxonomy[3] and ACM Computing Classification System[4] are used. The base CCF document is augmented with concept patterns present in other two reference documents. We have further augmented the term profile so obtained, by adding terms from a Computing dictionary. In the process of constructing this reference document, all terms occurring in the other two reference documents and the dictionary, are merged into appropriate categories of the CCF. All this taken together, resulted in a very useful CS domain concept list document which lists all major concepts and their subject area, and thus can be used for a specific kind of matching. This reduced the chances of missing any valid CS domain concept during the filtering process.

For identifying valid CS domain concept from the candidate list of concept patterns, we match every concept pattern in the concept list with the CS domain concept entries in the augmented ACM CCF document. A simple string matching, however, would not be useful and we used Jackard similarity measure to define a match. It is clear that using simple string matching will cause to miss a concept 'complexity of algorithm', if the reference document lists that concept as 'algorithm complexity'. The Jackard similarity measure is defined as:

$$Similarity(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where, **A** and **B** stand for the two concepts, $\boldsymbol{A} \cap \boldsymbol{B}$ is the set of common words in both concepts, $\boldsymbol{A} \cup \boldsymbol{B}$ is the set of union of words in both concepts and $|\boldsymbol{S}|$ stands for the number of elements in the set **S**. Here the concept denoted by **A** is the extracted concept as per the process described in 2.1 and the concept denoted by **B** refer to the concepts occurring in the augmented ACM CCF document. An important issue to decide in this process is what similarity score should be taken as a valid match. Setting it too high, would result in acceptance of only exact matches as valid CS domain concepts and setting it too low, will result in all candidate concepts with few common words being accepted as valid CS domain concept. Therefore, we used a moderate value of 0.5. Now, patterns **A** and **B** being, 'methods of numerical analysis' and 'numerical analysis methods', similarity score is 0.75, corresponding to pattern **A** being accepted as valid CS domain concept, despite different word order. We have also ranked the learning concepts in the order of their importance in a particular section/chapter, through a ranking scheme described in [7]. This allows us to a have only a selected number of important concepts being stored in the RDF structure and to provide for the information that may be required for a concept-based navigation system.

[2] http://ai.stanford.edu/users/sahami/CS2013/ironman-draft/cs2013-ironman-v1.0.pdf

[3] http://www.computer.org/portal/web/publications/acmtaxonomy

[4] http://www.acm.org/about/class/2012

3 Relation Extraction

While concepts as identified above give an indication of the subject matter discussed in a part of the eBook, it would be interesting to see which concepts are related and which concepts co-occur very frequently. This could be a measure of the set of concepts to be learned for understanding a particular topic. For identifying relations between concepts we used ReVerb [8], an open relation extractor which is known for its superior performance. The ReVerb relation extractor works on sentences and identifies those relation phrases which satisfy the syntactic and lexical constraints, and then finds a pair of Noun Phrase arguments for each identified relation. The ReVerb algorithm has three important differences with other relation extraction methods like TextRunner [9] and WOE [10]. The first difference is that here relation phrase is identified 'holistically' rather than word-by-word. Secondly, potential phrases are filtered based on statistics over a large corpus. And the third major difference is that ReVerb is 'relation first' rather than 'arguments first' relation extractor system.

The ReVerb system takes as input a POS-tagged and NP-chunked sentence and returns a set of (x; r; y) extraction triples. We already had POS tagged sentences available, hence applying ReVerb was easier. The ReVerb system is used to obtain relations in texts from sections/ chapters. However, not all the relations in this list refer to the relations between CS domain concepts. Therefore, our next step was to identify those relation pairs that express relationship between CS domain concepts. We have therefore used another level of filtering. We scan all the relation pairs and take a relation as a relation of interest if any of the two arguments that it has, is a valid CS domain learning concept in our system. This allows us to identify even those relations which add some meaning (by associating some other external concept) to a valid CS domain learning concept in the system. This helps in preparing both concept profile and concept relatedness profile of a section/ chapter of an eBook.

4 RDF Schema and Populating It

The list of valid CS domain concepts and relations between them, extracted from the eBook text, need to be stored in an appropriate data structure. The Resource Description Framework (RDF), the most frequently used framework now for semantic tagging and annotation of unstructured data, appears best for this purpose. RDF structures are essentially designed for reading by the machine [11]. In fact if the entire data on the World Wide Web, could be transformed to equivalent RDF structures and associated ontologies, we will have a much more useful and navigable Semantic Web. The task of storing all useful information from an eBook into an RDF structure is thus equivalent to doing semantic annotation/ tagging on it.

4.1 RDF Schema

The standard RDF syntax comprises of a set of triples, popularly visualized as a RDF graph. RDF triples contains three components namely subject, predicate and object. Consider the following RDF example:

```
<?xml version="1.0"?>
<RDF>
<Description about="http://www.w3schools.com/rdf">
<author>Jan</author>
<homepage>http://www.w3schools.com</homepage>
</Description>
</RDF>
```

In this, simple example the statement "The author of http://www.w3schools.com/rdf is Jan." is represented as a RDF structure. As could be seen from the example, the subject refers to resource and predicate refers to features of the resource. The predicate represents the relationship between the subject and object. In this example RDF representation the subject is "http://www.w3schools.com/rdf ", predicate is "author" and object is "Jan".

4.2 Populating the eBook RDF Schema

The information we obtain and extract from the eBook content is written programmatically into the RDF. The entire process from extraction of concepts and relations to writing it in RDF is automated and does not require any user intervention. We write three kinds of information into the eBook RDF. First of which is the concept and relation profile for each chapter/ section of the eBook as identified through concept and relation extraction system. Second kind of information is external metadata about the eBook collected from the World Wide Web such as reviews of the eBook, its rating, its price etc. Third information type deals with internally extracted and computed values about the eBook. This includes information such as author, number of pages etc. and the computed values such as its ACM CCF category, qualitative measures (coverage, readability and comprehensibility [1], [2]) and sentiment profile of the eBook. We used our old Sentiment Analysis program [12] and [13] for this purpose.

More precisely, the RDF contain rdfs:resources for eBook metadata, each chapter information and eBook review information. The eBook metadata comprises of eBook title, author, number of chapters, number of pages, corresponding google-eBook price, corresponding google-eBook rating, main and related category as per ACM CCF, its concept coverage score, its readability score and the computed sentiment polarity. All this information is stored in the RDF stricture for several purposes. First, we need a structure to store the information extracted from the eBook content. Secondly, we wish to store the useful information extracted in a machine-readable form that can serve as input to various application programs (such as tagging, concept-based navigation and recommendation systems).

5 Dataset and Experimental Results

We now present description of the dataset used and example results obtained at the different stages of the process. The algorithmic formulation takes a PDF eBook as

input and generates useful information and stores that into an RDF structure. We present here part of the intermediate and final RDF structure output results.

5.1 Dataset

We have performed our experimental evaluation on a moderate sized dataset collection of our own comprising of 30 eBooks on different subjects in Computer Science. Although our dataset comprises of eBooks in Computer Science, this algorithmic formulation can be extended to work on English language text in any domain.

5.2 Extracted Concepts

The POS tagged input sentences are used for probable concept identification and extraction, as described in section 2.1. We present here a small subset of concepts from the first chapter titled 'Introduction' of a popular eBook "Data Mining Concepts and Techniques". A total of 1443 concepts were obtained from it. Some of these concepts are:

```
Emphasis on mining, statistical methodology, information
repository, information networks, single predicate, natu-
ral language processing, intelligent decisions, data pro-
tection rights, optimization methods, business intelli-
gence.
```

5.3 Valid CS Domain Concepts

The concepts identified as above are then subjected to a CS domain filtering according to the process described in section 2.2. For the example result of 1443 probable concepts identified above, we get a reduced list of 96 CS domain concepts for the chapter titled 'Introduction' of the eBook "Data Mining Concepts and Techniques". Some of the concepts in the filtered list are:

```
business intelligence, knowledge management, entity rela-
tionship models, information technology, database manage-
ment system, semi supervised learning, cloud computing,
web search, query processing.
```

Here, the valid CS concepts displayed are first few valid learning concepts identified from a specific chapter of the eBook.

5.4 Relations Extracted

The next important output we obtain is the set of relations involving valid CS domain concepts. For the chapter titled 'Introduction' from the eBook "Data Mining Concepts

and Techniques", we obtained a total of 596 probable relations using the relation extractor. Out of this, the relations having both arguments as valid CS domain concepts are 62. Some of these relations so selected are:

(data mining tools, *provide*, data classification),
(statistical methods, *used to verify*, data mining results),
(search engines, *often need to use*, computer clouds),
(diverse data semantics, *poses challenges to*, data mining),
(data mining, *can be applied to*, any kind of data),
(data mining, *is the result of*, the evolution of machine),
(machine learning, *is highly related to*, data mining)

It would be important here to note that the relations are shown in italic to distinguish them from the arguments. Many of these relations are seen to add semantic meaning to the selected concepts and to some extent resemble the semantic network and slot and filler structures used in Artificial Intelligence. The relations identified are from a sentence boundary only and do not in any way refer to co-occurrence or co-reference probabilities of selected concepts in a part of an eBook. In a similar manner, we can identify relations involving just one valid CS domain concept.

For a better understanding and visualization of the identified relations, we can create/ generate a relation network from the relations so identified. This network visualization not only graphically illustrates the selected concepts, their relationship to other concepts and other key terms of the text; but also presents a cognitive map of a part of an eBook. We used Gephi[5] for drawing the network. An example relation network (with just 15 concepts) for the text from the chapter titled 'Introduction' of the "Data Mining Concepts and Techniques" eBook is shown in figure 1.

5.5 RDF Structure for the eBook

An example RDF structure for the eBook metadata for the eBook "Data Mining Concepts and Techniques" is as below.

```
<rdf:RDF
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:book="http://www.southasianuniversity.org/books/Data_Mining_Concepts_and_Techniques_Third_Edition#" >
<rdf:Description
rdf:about="http://www.southasianuniversity.org/books/Data_Mining_Concepts_and_Techniques_Third_Edition#metadata">
<book:btitle>Data Mining Concepts and Techniques Third Edition</book:btitle>
```

[5] https://gephi.org/

```
<book:author>JiaweiHan,MichelineKamber,Jian Pei</book:author>
<book:no_of_chapters>13</book:no_of_chapters>
<book:no_of_pages>740</book:no_of_pages>
<book:bconcepts>rule based classification, resolution, support vector
machines, machine learning,...</book:bconcepts>
<book:main_category>Intelligent Systems</book:main_category>
<book:main_cat_coverage_score>0.05110</book:main_cat_coverage_score>
<book:related_category>Programming fundamentals</book:related_category>
<book:related_category>Information Management</book:related_category>
<book:googleRating>User Rating: **** (3 rating(s))</book:googleRating>
<book:readability_score>56 (Fairly Difficult)
</book:readability_score>
</rdf:Description>
<book:related_category>Programmingfundamentals</book:related_category>
<book:related_category>InformationManagement</book:related_category>
<book:googleRating>User Rating: **** (3 rating(s))</book:googleRating>
<book:readability_score>56 (Fairly Difficult)
</book:readability_score>
</rdf:Description>
```

As can be seen from the RDF, the eBook metadata contains information like authors, number of pages, its computed ACM CCF category, price and rating etc. Further, it also stores some additional qualitative measures that we computed from the eBooks (as reported in [1], [2]). The RDF representation of a chapter node (chapter 6 here) of the RDF generated is as follows:

```
<rdf:Description
rdf:about="http://www.southasianuniversity.org/books/Data_Mining_Concept
s_and_Techniques_Third_Edition#chapter6">
<book:cconcepts>transaction database,datamining,hashtable,support sup-
port count,decisionanalysis,association rules...</ book:cconcepts>
<book:relation>the software display,may decide to purchase,a home secu-
rity system,0.3547584217041676</book:relation>
<book:relation>strong association rules,satisfy,both minimum sup-
port,0.44865456037003126</book:relation>
<book:relation>market basket analysis,may be performed on,the retail
data of customer transactions,0.9874720642457288</book:relation>
<book:relation>interesting association rules,was studied
by,Omiecinski,0.9628601549100813</book:relation>
.
</rdf:Description>
```

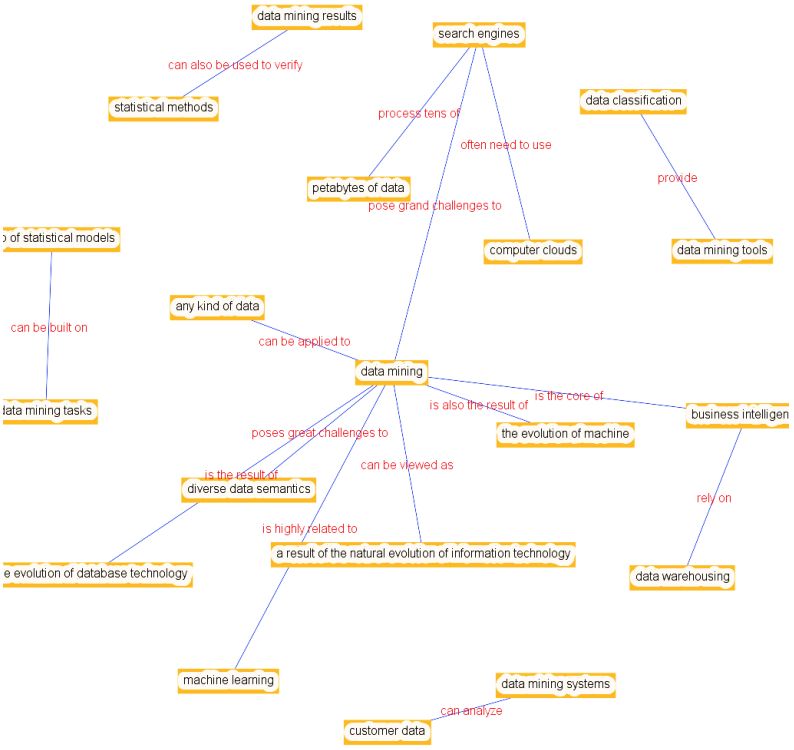


Fig. 1. Relation Network of Concepts

Similarly, the RDF graph for eBook metadata can be plotted. The RDF graph for the eBook used as running example is shown in the figure 2 below.

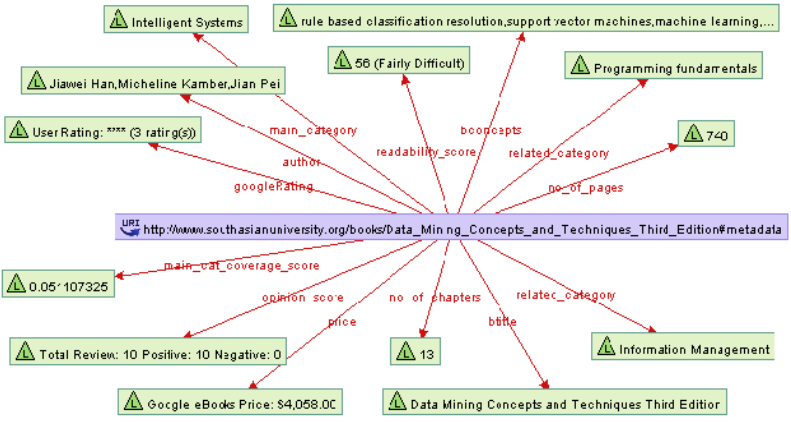


Fig. 2. RDF Graph for eBook Metadata

The RDF structure generated in this manner thus contains enormous amount of general and semantically useful information about every eBook. While it contains the general information about the eBook category, its author information, price, rating and reviews; it also contains chapter wise concept and relation profile for every eBook. The concept profile and the associated relation network provide useful information for semantic annotation of every part of the eBook [14]. Further, the relation network visualization can help the reader in better understanding of the learning concepts described in a particular part of the eBook.

6 Conclusion

In this paper, we have presented an algorithmic formulation for identifying valid CS domain learning concepts described in an unstructured text from eBook and relationships between the, through a linguistics-based approach. The system designed performs three kinds of computational tasks. First, it extracts various parts of an eBook and identifies important learning concepts and relations between concepts described in that part. This required sophisticated text parsing and mining. Secondly, the system automatically collects additional data about eBooks (such as reviews and ratings) for each eBook from the World Wide Web sources and performs computational tasks on them to draw useful inferences. This included sentiment analysis of eBook reviews to identify positive and negative reviews. Thirdly, some qualitative parameters such as category of an eBook and its readability and comprehensibility scores are also written to RDF structure.

The algorithmic formulation has possibilities of improvements in some aspects. One of them is POS tagging and identification of concepts and relations. With better POS tagger and relation extractor, we may be able to identify valid CS domain concepts in a better and more precise manner. The second possible improvement may be there in the relation network extraction and generation. We can compare the relation extraction using ReVerb with co-occurrence based concept-relatedness. The information extracted from eBook and then stored in RDF structure, makes it much easier to search from the eBook, navigate through one or a large collection of eBooks, and locate relevant chapter for obtaining knowledge about a learning concept. It can also help in designing a sophisticated eBook or concept learning recommender system. Further, since the reader can be presented with an auto-generated graphical visualization of the key concepts described in a part and how they are related to other concepts and entities, it becomes easier for the user to comprehend and also more effective for more retention of the learning that happened.

References

1. Relan, M., Khurana, S., Singh, V.K.: Qualitative Evaluation and Improvement Suggestions for eBooks using Text Analytics Algorithms. In: Proceedings of Second International Conference on Eco-friendly Computing and Communication Systems, Solan, India (2013)

2. Khurana, S., Relan, M., Singh, V.K.: A Text Analytics-based Approach to Compute Coverage, Readability and Comprehensibility of eBooks. In: Proceedings of the 6th International Conference on Contemporary Computing, Noida-India. IEEE Press (2013)
3. Justeson, J.S., Katz, S.M.: Technical terminology: Some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1) (1995)
4. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Data Mining for Improving Textbooks. ACM SIGKDD Explorations 13(2), 7–19 (2011)
5. Agrawal, R., Gollapudi, S., Kenthapadi, K., Srivastava, N., Velu, R.: Enriching textbooks through data mining. In: ACM DEV. (2010)
6. iText Open Source PDF Library for JAVA, http://www.api.itextpdf.com
7. Singh, V.K., Piryani, R., Uddin, A., Pinto, D.: A Content-based eResource Recommender System to augment eBook-based Learning. In: Proceedings of the 7th Multi-Disciplinary International Workshop in Artificial Intelligence, Krabi, Thailand. LNAI. Springer (2013)
8. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Conference on Empirical Methods in Natural Language Processing (2011)
9. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: International Joint Conference on Artificial Intelligence (2007)
10. Wu, F., Weld, D.S.: Open information extraction using Wikipedia. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 118–127. Association for Computational Linguistics, Morristown (2010)
11. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam.: Open Information Extraction: the Second Generation. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, pp. 3–10 (2011)
12. Singh, V.K., Piryani, R., Uddin, A., Waila, P.: Sentiment Analysis of Movie Reviews and Blog Posts: Evaluating SentiWordNet with different Linguistic Features and Scoring Schemes. In: Proceedings of 2013 IEEE International Advanced Computing Conference. IEEE Press, Ghaziabad (2013)
13. Singh, V.K., Piryani, R., Uddin, A., Waila, P.: Sentiment Analysis of Movie Reviews- A new feature-based Heuristic for Aspect-level Sentiment Classification. In: Proceedings of the 2013 International Muli-Conference on Automation, Communication, Computing, Control and Compressed Sensing, IEEE Press, Kerala (2013)
14. Uddin, A., Piryani, R., Singh, V.K.: Information and Relation Extraction for Semantic Annotation of eBook Texts. In: Thampi, S.M., Abraham, A., Pal, S.K., Rodriguez, J.M.C. (eds.) Recent Advances in Intelligent Informatics. AISC, vol. 235, pp. 215–226. Springer, Heidelberg (2014)

Mining for Marks: A Comparison of Classification Algorithms when Predicting Academic Performance to Identify "Students at Risk"

Lebogang Mashiloane and Mike Mchunu

School of Computer Science,University of the Witwatersrand, Johannesburg, South Africa
{lebogang.mashiloane,michael.mchunu}@wits.ac.za

Abstract. A major concern for higher education institutions is the high failure and drop-out rates amongst students, especially first year students. Tertiary institutions thus have a common interest in identifying students at risk of failing or dropping out. Previous research studies have identified factors that influence success/failure which include, but are not limited to, the students' personal information, academic background and social environment. This study aims to use the emerging field of Educational Data Mining as a preventative measure rather than reiterate factors that influence success. The first year student data collected and stored in the School of Computer Science at the University of the Witwatersrand has been utilised in this study. The study used the students' first semester/midyear mark to predict success/failure at the end of the academic year. This will assist in identifying students at risk of failing and could assist with early intervention. A modified version of the CRISP-DM methodology was used. The investigation was broken down into two phases: training and test phase. In the training phase, student data from the years 2009 to 2011 were modelled using the WEKA Explorer GUI. Three classifiers: J48 classifier, Naïve Bayes and Decision Table, were used for modelling and were also compared. Using both the run information from WEKA and performance metrics, the J48 classifier was shown to be the better performing algorithm in the training phase. This algorithm was then integrated into the back-end of the Success Or Failure Determiner (SOFD) tool, which was created specifically for this study. In the test phase 92% of the instances were predicted correctly. Furthermore 23 of the 25 students who failed were flagged. The research findings indicated that the midyear mark can be considered as a factor which correctly predicts the Computer Science I final year marks. After further investigation with larger sample sizes, the tool can be used practically in the school of Computer Science to identify students at risk of failing.

Keywords: Educational Data Mining, J48 Classifier, Decision Table, Naïve Bayes, WEKA, GUI.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 541–552, 2013.
© Springer International Publishing Switzerland 2013

1 Introduction

The identification of students at risk of failing is a major concern for educational institutions. This is interrelated to issues such as the high failure and drop-out rates which are common in tertiary institutions. Although most universities follow strict rules and processes to ensure that only the students most likely to succeed are selected (eg. Compulsory pre-requisites and minimum APS scores for different degrees), a larger portion of those students fall in the category of "Students at risk". Fraser and Killen ([7]) suggested that knowingly enrolling students into courses with knowledge that they have little chance of succeeding would be immoral; however this cannot always be guaranteed. A large group of institutions use the students' academic history as the deciding factor. An aggregate of the high school subjects is a more acceptable admission criteria [16]. Administrators and enrolment officers find themselves in a predicament with the lack of consistency within the literature available. However this continues to be the selection system used by many institutions. How can tertiary institutions then try to achieve the highest number of successful students from those that have been enrolled?

Even the students identified by the selection criteria and pre-enrolment tests as successful candidates may still experience barriers to learning or difficulties that could impact their academic performance. Factors that could affect the student include: their financial background, the social environment and at times the students expectations of the course. More specifically, first year students are said to usually have an inaccurate idea of what the course is and that could frustrate them [18]. Having the ability to classify students as being "at risk" will make it easier for the placing of resources where they are most needed and additionally for early intervention to occur. If there is a system in place for the prediction of final year results, students with the highest probability of failing will be identified. This would then raise a flag for lecturers, administrators, tutors and students before the commencement of final examinations.

Research has suggested that classification techniques can be used on educational data for the prediction of the academic performance of students [20]. Yadav and Pal([20]) used classification trees to identify students who are at risk of not performing well and try help them score better results, while Naik and Purohit([12]) used classification algorithms in-order to select the students with the greatest probability of being successful. This assisted in the placement of students into the course. These classification techniques, which were used for prediction, are part of a broader area called Educational Data Mining (EDM). This is an emerging field which focuses on obtaining knowledge from the large data sets in educational settings. The data is processed and transformed to find interesting patterns; discover similarities in characteristics and behaviour and present rules for forecasting.

Since all educational institutions store vast amounts of data about their students, this is a good place to start when trying to obtain knowledge about students or even lecturers. The stored data consists of the students' academic background, personal profile and the students' academic performance in that

institution. Most factors that could relate to success in educational institutions have been widely investigated. Since this study aims to create a tool that will identify "students at risk" before they fail, only factors available before examinations can be considered. Furthermore, with most studies looking at race, gender and academic history [19], this study will consider the use of the midyear mark as a predictor for final results. O'Bryne et al. ([13]) suggested that "Poor academic performance is often apparent early in a student's career". And although identifying "students at risk" could be helpful for all educational institutions, tertiary institutions in particular, face a greater challenge with students who drop out or fail in their first year of study.

In the School of Computer Science at the University of the Witwatersrand (Wits), to our knowledge, no tool has been created to predict the students' final mark. The aim of this study is to identify students who are at risk of failing or dropping out of their first year in Computer Science at Wits. This will be done by trying to answer the following questions:

- Which classification algorithm between the J48 Classifier, Decision Table and Naïve Bayes performs better in predicting the final mark from the first semester/midyear mark?
- Is the first semester/midyear mark a good factor for predicting the final Computer Science I mark?

2 Background

2.1 About the Computer Science I Class

The faculty of Science at Wits is home to the School of Computer Science where this research is based. In this study, the first year class in the School of Computer Science, also known as `COMS1000` or `CS-1`, was investigated. From the years 2010 to 2012, 391 students registered for the `CS-1` unit. The `COMS1000` unit consists of four different modules, namely: Basic Computer Organization (`BCO`), Data and Data Structures (`DDS`), Fundamental Algorithmic Concepts (`FAC`) and Limits of Computation (`LOC`).

Two modules are taught in each semester depending on the availability of lecturers and resources. In the years between 2010 and 2012, `BCO` and `DDS` were taught together in the first semester and FAC and LOC were taught in the second semester. Each module includes activities such as the attendance of lectures, tutorial and laboratory sessions, assessments and examinations. For each unit a decision code is used to show the academic performance of the student (Table 1). For the purpose of this research the unit decision codes were categorized into "Pass" and "Fail" for clarity on success or failure. Therefore, "PASS" and "PDS" fall under "Pass" while all other decision codes are considered as a "Fail"

2.2 Data Mining Techniques: Classification

Classification is one of the highly popular techniques in EDM. This is partly because of its association with the finding of unknowns and predicting of classes

Table 1. Unit Decision Codes

Decision Code	Description
PASS	Permitted to proceed
PDS	Pass with distinction
FAL	Failed to complete minimum requirements
FABS	Fail, absent from examination
FDF	Failed, deferred examination refused
FSB	Fail on sub-minimum
FNR	Fail, may not repeat unit
WDF	Deferred examination granted

or instances. It is an unsupervised technique in that models are trained and can thereafter predict unknown classes. The technique is divided into two phases: the training phase and the test phase.

- The training phase: data records with known classes are used to build a model.
- The test phase: the model built in the training phase is used to predict unknown classes

The three classification algorithms that were used in this study are the J48 Classifier, Decision Table and Naïve Bayes algorithm. These algorithms can be found in the Waikato Environment for Knowledge Analysis (WEKA) tool kit.

J48 Classifier. The J48 Classifier is one of the decision tree algorithms. It is the WEKA version of the C4.5 algorithm [6]. The J48 Classifier is suggested to be accurate, quick in the building of the model and has a simple presentation for the results[1]. Advantages of using the J48 Classifier or more specifically decision tree's is that they do not require a lot of data preparation and they present the results in a manner which is great for user understanding [4]. The algorithm involves the calculation of the information gain ratio for each attribute[20]. As shown in [20], the tree will be built using the maximum gain ratio as the root and pruning to remove unnecessary branches.

Decision Table. Decision tables result in a set of rules. Each rule has a condition and a conclusion. Decision tables are similar to decision trees but differ in that they produce rules instead of trees as output. Durant and Smith ([6]) suggest that Decision tables are simpler to use and the algorithm is less demanding on the computer. The cross validation of Decision tables is also incremental which makes the process faster [6]. The algorithm works by finding an attribute or combination of attributes which would best predict the class [6]. It uses forward selection which is a top to bottom search adding attributes as it goes [6].

Naïve Bayesian Network. The Naïve Bayes algorithm is based on the Bayes theorem which states (as shown in [15]):

$$P(A|B) = \frac{P(B|A)(P(A)}{P(B)} \tag{1}$$

P(A) being the prior probability of an event A and P(B) the prior probability of an event B. This would make $P(A|B)$ the conditional probability of A given B and similarly for $P(B|A)$ the conditional probability of B given A [15]. The algorithm is called Naïve because it assumes that attributes are independent and that there are no hidden attributes [14].The advantages of the algorithm include only requiring a small amount of data to make an estimation of the parameters [10]. Naïve Bayes can be represented as a network with connecting nodes.

2.3 Related Work

Little can be found in literature with regards to using the semester mark to predict the final year-end mark. A study similar to this one, looked at the academic performance of students in the first half of the semester [13]. O Bryne et al.[13] considered the relationship between the students' in-semester marks and their final semester mark rather than the relationship between the final semester marks with the final grade at the end of the year. Moreover, the study predicted the final exam marks and not the final year-end mark [13]. The final exam mark (predicted in this [13] study) and the final mark (predicted in the present study) differs in that the final exam mark does not include tests and other assessments. The conclusion that was made was that semester academic performance cannot on its own predict students at risk of failing. Although this study is comparable to the present investigation, [13] did not use classification techniques which will be used in the present study.

A study by [20] did not look at Computer Science students or midyear marks; however it used classification to predict the performance of engineering students. Three decision tree algorithms were used: C4.5, ID3 and CART algorithms. This encouraged the comparison of different algorithms. The aim of the study was to identify the 'weaker' performing students and assist in improving their results. Factors, such as the academic history of the student, the number of children in the family of the student and their admission type were considered. Results show that the C4.5 algorithm had the better accuracy of the three algorithms. The authors went on to suggest that algorithms such as the decision tree algorithms are effective in learning from previous year student data and then creating predictive models [20]. This was taken as a recommendation and one of the algorithms that were used in the present study was a decision tree.

3 Research Approach

3.1 Instruments

The two instruments which were used for this study are the Waikato Environment for Knowledge Analysis (WEKA) tool kit and the Success Or Failure

Determiner (SOFD) tool, which is a GUI based application created for this investigation. A modified version of the CRISP-DM was followed.

3.2 Research Design

Data Understanding. The data was collected from the School of Computer Science at Wits. It consisted of 391 first year students who have taken Computer Science as a major between the years 2010 and 2012. From these records, 257 would form part of the training data set (2010 and 2011) and the 134 would be the test data set (2012). The attributes in both data sets were the BCO and DDS mark, which are the two modules taught in the first semester of COMS1000 and the final Computer Science Decision code.

Data Processing. Most of the data collected from real world applications and databases is usually incomplete and contains some noise. In this case, the data was complete; all `BCO` and `DDS` marks had a corresponding final `COMS1000` grade attached to it. However, since the midyear mark was required, an average of the `BCO` and `DDS` mark was calculated and named the midyear mark. Thereafter the `BCO` and `DDS` marks were removed from the data sets. Therefore, after all the pre-processing, the data set that was left consisted of the 1st semester or midyear mark and the final `COMS1000` mark. The test data set was modified by removing the final `COMS1000` mark and placing question marks in the place where the predictions would need to be made. Both the training and test data sets were converted to .arff files, which is the preferred WEKA format.

Modelling. After the pre-processing phase, the data sets were ready for classification. The training data set was classified on the Explorer GUI of the WEKA tool kit using the J48 Classifier, Decision Table and Naïve Bayes algorithm. The Success Or Failure Determiner (SOFD) tool was created (Figure 1). This is a GUI based tool that has the WEKA General API embedded in it, so that the techniques from WEKA can be used. This GUI accepts a .arff file with the first semester results of all the students, uses the model built from the training data set and outputs the predicted final results for the students.

Evaluation and Algorithm Analysis. Performance metrics were used to analyze the performance of the three algorithms [9].

- Sensitivity: The true positives instances identified divided by the actual positive instances which are present.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

- Specificity: The true negatives identified divided by the actual negatives present.

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

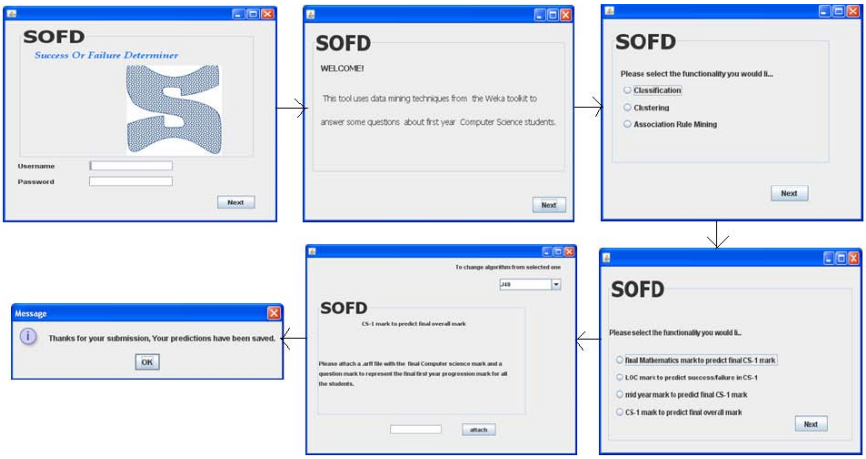


Fig. 1. The SOFD Tool

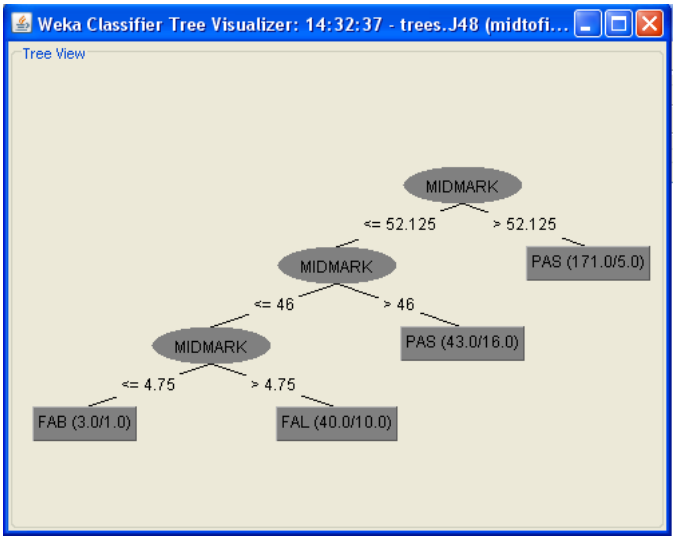


Fig. 2. Classification Tree from the J48 model

– Precision: The true positives instances identified divided by all the positives that have been identified.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

– Accuracy: The true negatives identified divided by all the negatives that have been identified.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

– False Positive Rate: Proportion of total number of negative instances classified as positive.

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

4 Results and Discussion

4.1 Comparison of Classifiers

Predictive models were obtained from the training data set using the J48 classifier, Decision Table and Naïve Bayes. The J48 classifier performed better with 86.38% of the instances being correctly identified in comparison to the 80.54% and 84.05% of the Decision Table and Naïve Bayes respectively (Table 2). The tree from the J48 Classifier is shown in Table 2. There was no difference in the time each algorithm took to build the models therefore that cannot be included in the deducing of the better performing algorithm. Furthermore, it was necessary to use performance measures to evaluate the performance of each algorithm. This highlighted the Decision Table as a good performer since it performed better in the specificity, precision and false positive rate as shown in Table 3. This meant that both the J48 classifier and the Decision Table were competitors for the better performing classification algorithm. For this reason, the run information from WEKA was investigated further. This tells us that the J48 Classifier has the highest Kappa statistic, at 0.5616, than the Decision Table and Naïve Bayes which had 0.4898 and 0.4581 respectively. This is an indication of the strength of the agreement between the two variables [11]. The Kappa values show that a classifiers performance is not due to chance and in this case from Table 4 we can say that the J48 classifier is of moderate strength. Furthermore the J48 Classifier had the lowest mean absolute error(0.0668) and root mean squared error(0.1919). Both these measures relate to the average error of the classifier. For these reasons the J48 Classifier is selected as the algorithm of choice for the

Table 2. Performance of Classifiers for 2010-2011

Algorithm	Time (s)	Correct (%)	Incorrect (%)
J48	0.00	**86.38**	**13.62**
Decision Table	0.03	73.61	26.39
Naïve Bayes	0.00	72.47	27.53

Table 3. Performance Measures

	J48	Decision Table	Naïve Bayes
Sensitivity	0.9602	0.9194	**0.9701**
Specificity	0.5714	**0.7321**	0.4643
Precision	0.8894	**0.9194**	0.8667
Accuracy	**0.8755**	0.8249	0.8599
False Positive Rate	0.4286	**0.2679**	0.5357

Table 4. Strength of agreement [11]

Value	Meaning
<0	poor
0-0.20	slight
0.21-0.40	fair
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect

prediction of the final COMS1000 grade from the 1st semester mark and used in the SOFD tool.

4.2 Prediction Using SOFD Tool

This investigation included using the 1st semester mark to predict the final `COMS1000` mark. Using the SOFD tool, the midyear mark test data set (2012) was classified using the J48 model built from the 2010-2011 (training) data set. The results are shown in Table 5.

Table 5. J48 Results

`J48 Classifier Results`	
Correctly Classified Instances:	92.5373%
Incorrectly Classified Instances:	7.4627%
Time taken to build model:	0.08 seconds
Kappa statistic:	0.777
Mean absolute error:	0.0361
Root mean squared error:	0.1218
Relative absolute error:	77.43.4504%
Root relative squared error:	60.8344%
Total Number of Instances:	134

The correlation between the 1st semester mark and the COMS1000 mark is shown in Figure 3. 92.54% of the test data was correctly classified. Therefore less than 8% of the 2012 CS-1 students had an incorrect prediction. From a data set of 134 students, 25 of these students failed Computer Science at the end of the year. 23 of these students were correctly predicted, this would mean not all the "students at risk" were flagged but an intervention could have occurred for 92% of these students. Bayer et al.[17] suggested that a method which tries to intervene with the dropping out of students, or in this case finding the "students at risk", should have minimum false negatives. When academic performance in

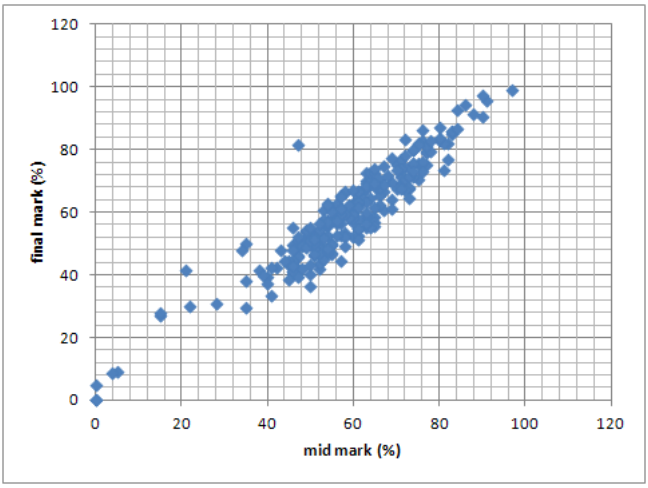


Fig. 3. Scatter Plot of 1st semester mark to COMS1000 mark

assessments in the first half of the first semester was compared with performance in the final examination, it was found that the correlation was poor and had no significance [13]. The difference with the study by O'Byrne et al. ([13]) and this present study is that we look at the final first semester mark and not correlation with assessments within the semester. Additionally, the present study uses classification for prediction purposes and a scatter plot just to illustrate the significance of the correlation.

5 Conclusion and Future Work

There are three major findings in the present study. Since 92.54% of the students results were correctly predicted, the midyear mark can be presented as an effective factor for the prediction of first year Computer Science final results. Although previous evaluations of the midyear and final year mark showed little relation between the two, those studies did not involve classification techniques. This gives us a new method to tackle the influence of the midyear mark on the final year mark. The prediction of the final mark will assist in the identifying of students who are "at risk" of failing and therefore, allow for early intervention to occur. Furthermore, students, administrators and lecturers will benefit greatly from this.

In the comparison of the three algorithms, the J48 Classifier was found to be the better performing algorithm when compared to the Decision Table and Naïve Bayes. This requires further investigation since in some cases the Decision Table was shown to be a good competitor for the title. This opens up discussions into which classification algorithms perform better specifically in the classifying of educational data.

And finally, the SOFD tool, which was a product of this investigation, will not only allow for further research into mining from educational data at Wits but will also assist the School of Computer Science in predicting marks and identifying students at risk before they write their final examinations.

This study contributes to Educational Data Mining by highlighting the J48 Classifier as one of the better algorithms for predicting academic performance. Furthermore, it contributes to education by not only helping lecturers and administrators identify "students at risk" but by also allowing for intervention for the students before they drop out or fail.

Future work from this study will include an analysis of other factors that could assist in the predicting of student results such as gender, race and course combinations. Furthermore, other data mining techniques including clustering and association rule mining will be investigated in the area of obtaining knowledge for educational data sets. Knowledge obtained for all these investigations will be integrated into the SOFD tool. More modifications will allow for the different schools in the university to be able to predict the students' final grades and identify those students who could potentially drop out or fail using the SOFD tool.

Acknowledgments. I would like to thank my supervisor Mr. Mike Mchunu for his assistance and guidance. I would also like to thank my husband, Landi Mashiloane, and family for their prayers and continuous support.

References

1. Bhullar, M.S., Kaur, A.: Use of Data Mining in Education Sector. Lecture Notes in Engineering and Computer Science, vol. 2200, pp. 513–516 (2012)
2. Butcher, D.F., Muth, W.A.: Perdicting performance in an introductory computer science course. ACM 28(3) (1985)
3. Campbell, P., McCabe, G.: Predicting the success of freshmen in a computer science major. Commun. ACM 27(11), 1108–1113 (1984), http://doi.acm.org/10.1145/1968.358288
4. Chandra, E., Nandhini, K.: Predicting student performance using classification techniques. In: Proceedings of SPIT-IEEE Clloquium and International Conference
5. Delavari, N., Phon-Amnuaisuk, S., Beikzadeh, M.: Data mining application in higher learning institutions. International Journal of Informatics in Education 7(1), 31–54 (2008)
6. Durant, K.T., Smith, M.D.: Predicting unix commands using decision tables and decision trees. In: Proceedings of the Third International Conference on Data Mining, pp. 427–436 (September 2004)
7. Fraser, W.J., Killen, R.: Factors influencing academic success or failure of first-year and senior university students: do education students and lecturers perceive things differently. South African Journal of Education 23(4), 254–260
8. Garcia-Saiz, D., Zorrilla, M.: Comparing classification methods for predicting distance students performance. In: JMLR: Workshop and Conference Proceedings 17, 2nd Workshop on Applications of Pattern Analysis 2011, pp. 26–32 (2011)

9. Kumar, V., Rathee, N.: Knowledge discovery from database using an integration of clustering and classification. IJACSA - International Journal of Advanced Computer Science and Applications 2(3), 29–33 (2011)
10. Panday, U.K., Pal, S.: Data Mining: A prediction of performer or underperformer using classification. International Journal of Computer Science and Information Technologies 2, 686–690 (2011)
11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977), http://www.jstor.org/stable/2529310
12. Naik, N., Purohit, S.: Article: Prediction of final result and placement of students using classification algorithm. International Journal of Computer Applications 56(12), 35–40 (2012), published by Foundation of Computer Science, New York, USA
13. O'Byrne, J., Britton, S., George, A., Franklin, S., Frey, A.: Using academic predictors to identify first year science students at risk of failing. CAL-laborate International 17 (2009)
14. Osmanbegović, E., Suljić, M.: Data mining approach for predicting student performance. Economic Review 10(1) (2012)
15. Riesenfeld, R.: Bayes' Theorem (2011), http://www.eng.utah.edu/~cs5961/Resources/bayes.pdf
16. Rauchas, S., Rosman, B., Konidaris, G.: Language performance at high school and success in first year computer science. SIGCSE 2006 (2006)
17. Obsivac, T., Popelinsky, L., Bydzovska, J.B.J.G.,, H.: Predicting drop-out from social behaviour of students, p. 103
18. Turner, E.H., Turner, R.M.: Teaching entering students to think like computer scientists. SIGCSE (2005)
19. Wimshurst, K.J., Wortley, R.K.: Academic success and failure: Student characteristics and broader implications for research in higher education. In: Effective Teaching and Learning. Griffith Institute for Higher Education (2005)
20. Yadav, S., Pal, S.: Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal (WCSIT) 2(2), 51–56 (2012)

Determining Students Expectation in Present Education System Using Fuzzy Analytic Hierarchy Process

S. Rajaprakash[1] and R. Ponnusamy[2]

[1] Department of Computer Science and Engineering, Aarupadai Veedu Institute of Technology,Chennai, India
srajaprakash_04@yahoo.com
[2] Deparement of Computer Science and Engineering , Madha Engineering College, Chennai
r_ponnusamy@hotmail.com

Abstract. This work is based on the extensive analysis of the student expectation from the Present Education System. It is a major focused area for the bright future of the country . Selection of the right educational institution for the student is a crucial decision for the parents, this requires optimizing a number of criteria based on their expectations. .Education the Engineering Colleges in Tamilnadu is considered as sample with Fuzzy Analytic Hierarchy Process (FAHP). FAHP is applied to find the degree of each criterion of the education institutions. In this study nine kinds of key attributes are used based on expert opinion teachers and academicians of Technical educational institutions. Using these attributes a comparison matrix and triangular membership function have been formed which helps to evaluate the attributes of the present Technical institutions in Tamilnadu.

1 Introduction

Present Education system in Tamil Nadu (India) has lot of options from students perspective. It becomes complex to select the future course of their education, that too in which Institution Tamil Nadu government has given permission to start lot of Education Institution for the welfare of the student's future. So in Tamil Nadu lot of Engineering, Medical, Arts and Science, Law Institution ware started in the past 12 years. This scenario was different before 12 years where only 220 Engineering colleges were available. But now in 2013 more than 600 Engineering Colleges are available and in the Arts and Science more than 800 Colleges are available . Apart from that more than 20 Deemed universities are available[1].

In the last academic year 8.53 lakhs student passed their +2 exams and they are ready for higher Education. Though around 2.6 lakhs Engineering seats were available they were not fully filled for the past 7 years.This has created a situation, where, there are lots of options for the student to select the best college for their future. In the last year the total Capacity of Engineering seats was 2, 62,164 across

[1] http://www.icbse.com/universities/deemed#tamilnadu, Last accessed July 30-10-2013

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 553–566, 2013.
© Springer International Publishing Switzerland 2013

Tamil Nadu. Out of which only 1, 82,493 were filled during the last academic year. still about 79000 engineering seats remained unfilled in the last academic year[2].

From the Institutional point of view it is very difficult to fill all available and hence to survive. In the last year more than 100 Engineering colleges were planning to close. To overcome this critical situation of the Technicals Institution especially in the private sector there is a need those to upgrade those colleges or universities to meet the students' expectation and to follow to Government of India guidelines.

Let us try ot undersatand how the student is selecting the Engineering College. From the whether it is student's or the parent's point of view?. The selection of the Engineering colleges in Tamil Nadu has always been considered a major problem of uncertainty. The cost of Engineering education very high. Hence the selection the engineering College becomes a crucial step for student's and parents. Selecting a right Engineering College to meet their future career or organizational requirement is a difficult task. In needs a full examination of various factors involved. A well prepared questionnaire is used for determining the user preferences for this process.

There are several methods of multiple-criteria decision making (MDCD) methods to evaluate several alternatives to achieve a certain goal. AHP methods is used to solve many complex decision problems. The AHP (analytic hierarchy process) is one of the good methods. In the present work we are using FAHP (Fuzzy Analytic hierarchy process) is used to get more accurate results.

2 Literature Review

In the real time problem, it is very difficult to extract the correct data of input and output and tackle them with crisp number which will reflect human's appraisals related to pairwise comparison. AHP approach as decision making was proposed by Saaty [1]. Using the AHP some engineering applications (complex problem) are solved by Evangelos Triantapbhyllou et.al[2].Fuzzy Analytical processes have been used during this process to determine the importance of weights of imprecise ranking of customer requirements by c. K.. Kwong et.al [3].With FAHP approach selection of the aesthetic attributes of car profile and their relative importance by H.C. Yadav et.al[4].Selection of the correct DBMS for software using Fuzzy Analytical processes. The DBMS alternatives are evaluated by assigning a rating scale by F. Ozgur Catak et.al [5]. In the regenerative technology using the Fuzzy Delphi Method obtain the critical factors and using FAHP provide the systematic approach towards the technology selection by Cheng-Haw et.al[6]. Smaller problem were using the AHP and its various real time application were studied by Satty.T.L[7]. How the key factors affect success in E-commerce using fuzzy AHP and its help researched and managers to determine the drawbacks and opportunities by Feng Kong et.al [8]. In 2005 a frame work was developed for an intelligent risk management system based on the Australia and New Zealand Risk Management Standard using AHP and Bayesian Belief Networks by A.Ahmed et.al[9]. In 2004 a New product Development environment using the Fuzzy AHP. In this a real-life manufacturing system us used

[2] http://www.thehindubusinessline.com/news/states
Last accessed July 30-10-2013

and tested by simulation analysis and integrated with the fuzzy AHP method by Zeki Aya [10]. Using the Fuzzy Delphi Method and Fuzzy analytic Hierarchy process is applied And determine the critical factors of the regenerative technologies and find the importance degree of each criterion as the measurable indices of the regenerative technologies by Yu-Lung Hsu et.al [11]. Influencing factors of online consumer purchasing behaviors by using the survey and AHP by LI Guo [12]. With FAHP for evaluating suitable method of selection to economic cocoon traits development is silkworm breeding by Meysam Shaverdi et.al[13]. A new frame work formed which is Using the rough sets and AHP select the best supplier using the several criteria by Berkir et.al[14]. In 2012 a new framework compared with FAHP and Rough set for predicting highest and lowest temperature by BP neural network for abnormal weather alerts by Dan Wang et.al[15]. (2012) Using the fuzzy Delphi Method and AHP and Fuzzy AHP construct the hierarchy to evaluate hotel atmosphere and incorporates a FAHP to integrate the knowledge and finding the priorities of criteria by Yen-Chen et.al [16]. (2012) In the supply chain management selection the vendor is a complex one. Using Fuzzy AHP vendor selection problems evaluated by Saroj Koul et.al [17]. Construction of membership functions and the importance of the criteria for selecting a bank account for students was evaluated by Alessio Ishizaka et.al [18]. In 2013 strategic energy technology roadmap in the area of energy technologies against high oil prices was studied by Seong Kon Lee et.al [19]. In the supply chain Management using the FAHP the consistency approaches by factor analysis that determines the adoption and implementation of green supply chain management in Indian Pharmaceutical companies has been done by Ajay Verma et.al [20]. (2013)Using the hybrid fuzzy AHP model solve the location choice problem of international distribution centre were started in the global logistics of multinational corporation by Chien-Chang et.al [21]. ERP (Enterprise Resource planning) is a software packages are used by many industries planned aspiration level is not up the level. In (2013) FAHP and DEMATEL to evaluate ERP critical success Factor in the industries by Saeed Rouhani et.al [22]. A new methodology is used to improve the quality of prioritization of an employee's performance measurement attributes under fuzziness using Extent Fuzzy Analytic Hierarchy Process by Aggarwal et.al[23]. (2012) In the phamaceutial industry prioritizing the criteria involved in selection of global supplier using the FAHP by Aysegul Tas et.al [24]. Use of the FAHP was utilized to identify the main risks then Interpretive Structural Modelling (ISM) was used to illustrate the interrelationship of those risk mitigations in the development of the mangosteen supply chain in Indonesia by Rentno et.al [25].

In the literature reviewed above many researchers proposed a lot of fuzzy and AHP approaches in the present work fuzzy AHP model is used for dealing with the situation of how the students are selecting the colleges in Tamilnadu based on Nine Major criteria in the decision-making process

3 Fuzzy Set Theory

The fuzzy sets have been introduced by Lotfi A. Zadeh in 1965. Fuzzy sets are an extension of classical set theory and are used in fuzzy logic. In crisp set theory the membership of elements in relation to a set is assessed in binary terms according to a

crisp condition but fuzzy set theory allows the gradual assessment of the membership of elements in relations to a set, this is described with the aid of a membership function valued in the real in [0,1]. The membership function maps crisp elements in the universe of discourse to elements' degree of membership with a certain interval, which is usually [0,1]. Then , the degree of membership specifies the extent to which a given element belongs to a set or is related to a concept. A fuzzy number is a special fuzzy set $F = \{x, \mu_F(x)), x \in R\}$ where x are values on the real line, R: $-\infty < x, +\infty$ and $\mu(x)$ is a continuous mapping from R to the closed interval [0,1]. A triangular member ship in fuzzy set theory is

$$\mu_F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ \dfrac{c-x}{c-b} & b \le x \le c \\ 0 & x > c \end{cases} \tag{1}$$

And the triangular fuzzy number can be characterized by

$$\tilde{M} = [a^{\alpha}, c^{\alpha}] = [(b-a)\alpha - (c-b)\alpha + c] \forall \alpha \in [0,1]. \tag{2}$$

$$\tilde{M} \oplus \tilde{N} = [M_L^{\alpha} + n_L^{\alpha}, m_R^{\alpha} + n_R^{\alpha}]$$
$$\tilde{M} \Theta \tilde{N} = [M_L^{\alpha} - n_L^{\alpha}, m_R^{\alpha} - n_R^{\alpha}]$$
$$\tilde{M} \otimes \tilde{N} = [M_L^{\alpha} n_L^{\alpha}, m_R^{\alpha} n_R^{\alpha}]$$
$$\tilde{M} \Xi \tilde{N} = [M_L^{\alpha} / n_L^{\alpha}, m_R^{\alpha} / n_R^{\alpha}]$$

In this work , triangular fuzzy number , $\tilde{1}$ to $\tilde{9}$ pairwise used to represent subjective pair wise comparisons. Here triangular fuzzy numbers are defined according to that fuzzy numbers are defined with the corresponding membership function given below.(Fig-1)

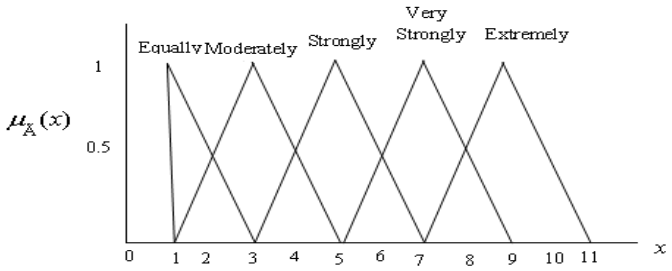


Fig. 1. Triangular fuzzy number of $\tilde{1}$ $\tilde{3}$ $\tilde{5}$ $\tilde{7}$ $\tilde{9}$

Here the α-cut and μ into the fuzzy AHP matrix taking care of the accuracy of the measurement . The α-cut value depends upon the decision maker or expert's confidence level of the judgments. Here the α-cut =0. 5 and μ=0. 5 using this we show the below diagram for (6,7,8) in Fig-2

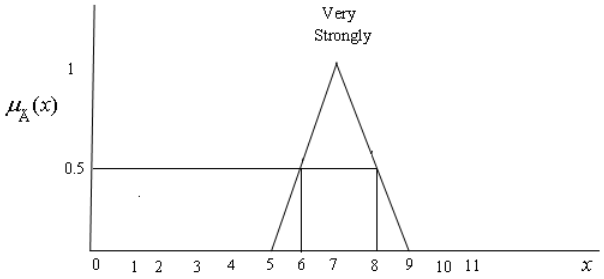


Fig. 2. Membership function with α -cut

3.1 Calculating Importance Weights of Fuzzy Numbers

The lower limit and upper limit of the fuzzy number with respect to the alpha can be defined by equation (2)

$$\begin{aligned}
&\tilde{1} = [1,3-2\alpha] \\
&\tilde{3} = [1+2\alpha,5-2\alpha] \qquad \tilde{3}^{-1} = \left[\frac{1}{5-2\alpha},\frac{1}{3+2\alpha}\right] \\
&\tilde{5} = [3+2\alpha,7-2\alpha] \qquad \tilde{5}^{-1} = \left[\frac{1}{7-2\alpha},\frac{1}{3+2\alpha}\right] \\
&\tilde{7} = [5+2\alpha,9-2\alpha] \qquad \tilde{7}^{-1} = \left[\frac{1}{9-2\alpha},\frac{1}{5+2\alpha}\right] \\
&\tilde{9} = [8+2\alpha,11-2\alpha] \qquad \tilde{9}^{-1} = \left[\frac{1}{11-2\alpha},\frac{1}{7+2\alpha}\right]
\end{aligned} \tag{3}$$

4 Analytic Hierarchy Process (AHP)

The AHP which is a powerful tool in applying Multicriteria Decision Analysis was introduced by Satty in 1980 [4]. The Analytic Hierarchy Process is a powerful and flexible decision making technique to help with the decision for both qualitative and quantitative aspects . In this method, finding the weights or priority vector of the alternatives or the criteria is required. For this purpose pair wise comparison matrix developed.

Table 1. Analytic Hierarchy Process pair wise comparison scale

Numerical values	Linguistics scale	Explanation
1	Equal importance of both elements	Two elements contribute equally
3	Moderate importance of one element over another	Experience and judgment favour one element over another
5	Strong importance of one element over another	An element is very strongly dominant
7	Very strong importance of one element over another	An element is favoured by at least an order of magnitude
9	Extreme importance of one element over another	An element is favored by at least more than an order of magnitude
2,4,6,8	Intermediate values	Used to compromise between two judgments

5 Fuzzy AHP

Fuzzy analytic Hierarchy Process proposed by Laarhoven and Pedrycz (1983) which is an application of the combination of Analytic Hierarchy Process (AHP) and Fuzzy Theory. In FHAP converts the opinions of experts from previous definite values to fuzzy numbers and membership functions.

Table 2.

Linguistic Scale	Triangular Fuzzy scale	Fuzzy Number	Intensity of importance
Equally important	(1,1,1)	$\tilde{1}$	1
Moderately more important	(1,3,5)	$\tilde{3}$	3
Strongly more important	(3,5,7)	$\tilde{5}$	5
Very strongly more important	(5,7,9)	$\tilde{7}$	7
Extremely more important	(7,9,11)	$\tilde{9}$	9

5.1 Methodology in Fuzzy AHP

FAHP proves that many concepts in real time have fuzziness. So the opinions of decision makers are converted from previous definite values to0 fuzzy numbers in the FAHP [5].

The following step is involved in FAHP method

1. Using the questionnaires collects opinions from the expert by using the linguistic variable in questionnaires.
2. The triangular fuzzy number is calculated from the factor given by experts. Using the geometric mean proposed by Klir and Yuan (1985) find out the significance triangular fuzzy number of the alternate factor is found.

3. Constructing the fuzzy comparison matrix by using a triangular fuzzy number via pairwise comparison, the fuzzy judgment matrix $\widetilde{A}(a_{ij})$ is constructed as given below

$$\tilde{A}_{ij} = \begin{bmatrix} 1 & \tilde{a}_{12} & \tilde{a}_{13} & \dots & \tilde{a}_{1(n-1)} & \tilde{a}_{1n} \\ \tilde{a}_{21} & 1 & \tilde{a}_{23} & \dots & \tilde{a}_{2(n-1)} & \tilde{a}_{2n} \\ & & 1 & \dots & & \\ & & & \dots & & \\ \tilde{a}_{(n-1)1} & \tilde{a}_{(n-1)2} & \tilde{a}_{(n-1)3} & \dots & 1 & \tilde{a}_{(n-1)n} \\ \tilde{a}_{n1} & \tilde{a}_{n2} & \tilde{a}_{n3} & \dots & \tilde{a}_{n(n-1)} & 1 \end{bmatrix}$$

Where $a_{ij} = \begin{cases} \tilde{1} \\ 1 \end{cases}, \tilde{3}, \tilde{5}, \tilde{7}, \tilde{9} \ \ or \ \tilde{1}^{-1}, \tilde{3}^{-1}, \tilde{5}^{-1}, \tilde{7}^{-1}, \tilde{9}^{-1} \ , i = j, i \neq j$

4. Finding the fuzzy Eigen values. A fuzzy Eigen value , $\tilde{\lambda}$, is fuzzy number solution to

$$\tilde{A}\tilde{x}=\tilde{\lambda}\tilde{x} \tag{4}$$

Where $\tilde{A}$ nxn fuzzy matrix containing fuzzy numbers $\tilde{a}_{ij}$ and $\tilde{x}$ is a non zero nx1 fuzzy vector containing fuzzy number $\tilde{x}_i$. In fuzzy multiplications and additions using the interval arithmetic and α cut, Equation 1 is equivalent to

$$[a^{\alpha}_{i1l}x^{\alpha}_{1l}, a^{\alpha}_{i1u}x^{\alpha}_{1u}] \oplus \dots \oplus [a^{\alpha}_{i1l}x^{\alpha}_{nl}, a^{\alpha}_{inu}x^{\alpha}_{nu}] = [\lambda x^{\alpha}_{1l}, \lambda x^{\alpha}_{iu}]$$

Where

$$\tilde{A}=[a_{ij}] \quad , \ \tilde{x}^t = (\tilde{x}_1, \dots, \tilde{x}_n)$$

$$\tilde{a}^{\alpha}_{ij} = [a^{\alpha}_{ijl}, a^{\alpha}_{i1u}] \quad \tilde{x}^{\alpha}_1 = [x^{\alpha}_{il}, a^{\alpha}_{ju}], \tilde{\lambda}^{\alpha} = [\lambda^{\alpha}_1, \lambda^{\alpha}_u]$$

$$\text{where } 0 < \alpha \leq 1$$

5. Degree of satisfaction of the judgment matrix $\tilde{A}$ is estimated by the index of optimism μ. The larger value of the index μ indicates the higher degree of optimism. The index optimism is a linear convex combination defined as

$$\tilde{a}^{\alpha}_{ij} = \mu\tilde{a}^{\alpha}_{iju} + (1-\mu)a^{\alpha}_{ij1} \quad \forall\mu \in [0,1] \tag{5}$$

While μ is fixed, following crisp judgment matrix can be obtained after setting the index of optimism μ, in order to estimate the degree of satisfaction.

$$\tilde{A}=\begin{bmatrix} 1 & \hat{a}_{12}^{\alpha} & \dots & \hat{a}_{1n}^{\alpha} \\ \hat{a}_{21}^{\alpha} & 1 & \dots & \hat{a}_{2n}^{\alpha} \\ \dots & \dots & \dots & . \\ \hat{a}_{n1}^{\alpha} & \hat{a}_{n2}^{\alpha} & \dots & \end{bmatrix}$$

The eigenvector is calculated by fixing the μ value and identifying the maximal Eigen value.

6. Determining the weights of attributes

According to Saaty (1980) weighting vector for each pair wise matrix. the eigenvector is calculated by fixing the μ value andEigen valueg the maximal Eigen valus or The Eigen value and Eigen vector are calculated[6].

By Normalization of both the matrix of paired comparisons and evolution of priority weights $\lambda_{\max}$ is calculated. To control the results of this method, the consistency ratio for each matrix are calculated. The consistency is expressed by the following equation.

$$\text{Consistency index(CI)} = \frac{\lambda_{\max} - n}{n-1} \tag{6}$$

The consistency ratio is used to estimate the consistency of pairwise comparisons

Where RI is selected from the number of comparisons from the following Table 3

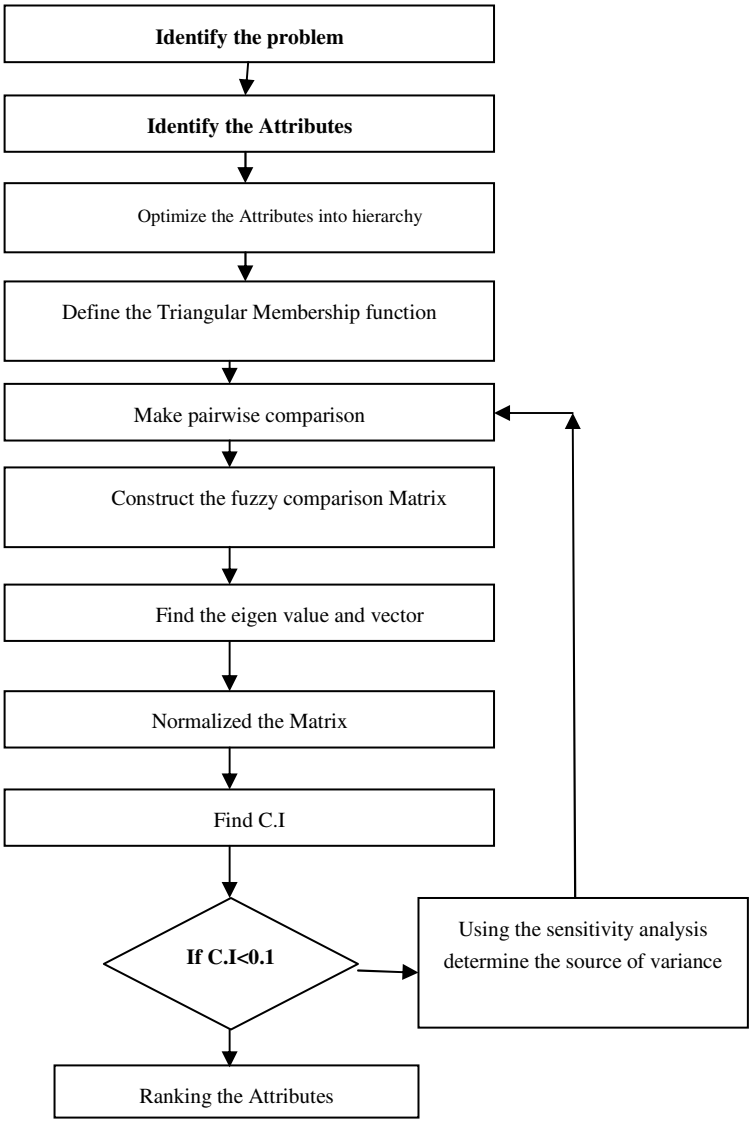
Table 3.

Matrix Rank	2	3	4	5	6	7	8	9	10
RI	0.00	0.58	0.90	1.12	1.24	1.35	1.41	1.45	1.49

$$\text{Consistency ratio (CR)} = \frac{CI}{RI} \tag{7}$$

If CR<0.1 then acceptable else the original values in the pairwise comparison matrix must be revised by the decision maker.

5.2 Fuzzy AHP Flow Chart



6 Attributes of Educational Institution

In order to evaluate different students' expectation from the educational institutions,we have developed a simplified decision model. This simplification redces the complexity of the decision. In this study important factor is identified and categories into Nine attributes from the experts from the educational institution. The attributes are

A1-Infrastructure
A2-Faculty Standard
A3-Accessiblity from city
A4-Hotel and food
A5-Discipline
A6-Bus facility
A7-Placement carrier
A8- Innovative ideas
A9- Fee structure of the institution.

After finalizing the assessment of relative importance of the attributes of the educational institution, the fuzzy comparison matrix formed from Table 1

Table 4. Fuzzy comparison matrix from the educational institution attributes

$$\text{FCM} = \begin{bmatrix} 1 & \tilde{3}^{-1} & \tilde{3} & \tilde{3}^{-1} & \tilde{3}^{-1} & \tilde{3} & \tilde{3}^{-1} & \tilde{1} & \tilde{3} \\ \tilde{3} & 1 & \tilde{5} & \tilde{3} & \tilde{3} & \tilde{3} & \tilde{3}^{-1} & \tilde{3} & \tilde{3} \\ \tilde{3}^{-1} & \tilde{5}^{-1} & 1 & \tilde{5}^{-1} & \tilde{5}^{-1} & \tilde{3} & \tilde{5}^{-1} & \tilde{3}^{-1} & \tilde{3}^{-1} \\ \tilde{3} & \tilde{3}^{-1} & \tilde{5} & 1 & \tilde{1} & \tilde{3} & \tilde{5}^{-1} & \tilde{3} & \tilde{3}^{-1} \\ \tilde{3} & \tilde{3}^{-1} & \tilde{5} & \tilde{1} & 1 & \tilde{7} & \tilde{9}^{-1} & \tilde{5}^{-1} & \tilde{5}^{-1} \\ \tilde{3}^{-1} & \tilde{3}^{-1} & \tilde{1} & \tilde{3}^{-1} & \tilde{7}^{-1} & 1 & \tilde{5}^{-1} & \tilde{3}^{-1} & \tilde{3}^{-1} \\ \tilde{3} & \tilde{3} & \tilde{5} & \tilde{5} & \tilde{9} & \tilde{5}^{-1} & 1 & \tilde{3} & \tilde{3} \\ \tilde{1} & \tilde{3}^{-1} & \tilde{3} & \tilde{3}^{-1} & \tilde{5} & \tilde{3} & \tilde{3}^{-1} & 1 & \tilde{1} \\ \tilde{3}^{-1} & \tilde{3}^{-1} & \tilde{3} & \tilde{3} & \tilde{5} & \tilde{3} & \tilde{3}^{-1} & \tilde{1} & 1 \end{bmatrix}$$

After finalizing the fuzzy comparison matrix by the experts, the triangular membership function and α-cuts=0.5 and μ=0.5 are used in this study, substituting the value of μ=0.5 and α=0.5 in the equation (3) we get the following Fuzzy comparison Matrix .

$$FCM = \begin{bmatrix} 1 & \left[\frac{1}{4},\frac{1}{2}\right] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{4},\frac{1}{2}\right] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & [1,2] & [2,4] \\ [4,6] & 1 & [4,6] & [4,6] & [2,4] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & [2,4] & [2,4] \\ \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{6},\frac{1}{4}\right] & 1 & \left[\frac{1}{6},\frac{1}{4}\right] & \left[\frac{1}{6},\frac{1}{4}\right] & [1,2] & \left[\frac{1}{6},\frac{1}{4}\right] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{4},\frac{1}{2}\right] \\ [2,4] & \left[\frac{1}{6},\frac{1}{4}\right] & [4,6] & 1 & [1.2] & [2,4] & \left[\frac{1}{6},\frac{1}{4}\right] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] \\ [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & [4,6] & \left[\frac{1}{2},1\right] & 1 & [1,3] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{3}{2},\frac{5}{2}\right] & [1,2] \\ \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{2},1\right] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{3}{2},\frac{5}{2}\right] & 1 & \left[\frac{1}{4},\frac{1}{6}\right] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{4},\frac{1}{2}\right] \\ [2,4] & [2,4] & [4,6] & [4,6] & [2,4] & [4,6] & 1 & [2,4] & [2,4] \\ \left[\frac{1}{2},1\right] & \left[\frac{1}{4},\frac{1}{2}\right] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & [1,3] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & 1 & [1,2] \\ \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{4},\frac{1}{2}\right] & [2,4] & [2,4] & \left[\frac{1}{2},1\right] & [2,4] & \left[\frac{1}{4},\frac{1}{2}\right] & \left[\frac{1}{2},1\right] & 1 \end{bmatrix}$$

6.1 Degree of Optimization

The highest value of the index μ gives the highest degree of optimization. And the index of optimum is a linear convex defined by equation (5). The following crisp decision

matrix can be obtained from the index of optimism value μ. Here μ=0. 5 is used to get fuzzy comparison matrix into a crisp comparison matrix. So the crisp comparison matrix (CCM) obtained after fixing the value of μ=0. 5 in the equation (5).

$$CCM=\begin{bmatrix} 1 & 0.375 & 3 & 0.375 & 0.375 & 3 & 0.375 & 1.5 & 3 \\ 5 & 1 & 5 & 5 & 3 & 3 & 0.375 & 3 & 3 \\ 0.208 & 0.208 & 1 & 0.208 & 0.208 & 1.5 & 0.208 & 0.375 & 0,375 \\ 3 & 0.208 & 5 & 1 & 1.5 & 3 & 0.208 & 3 & 0.375 \\ 3 & 0.375 & 5 & 0.666 & 1 & 2 & 0.375 & 2 & 1.5 \\ 0.375 & 0.375 & 0.666 & 0.375 & 0.5 & 1 & 0.208 & 0.375 & 0.375 \\ 3 & 3 & 5 & 5 & 3 & 5 & 1 & 3 & 3 \\ 0.666 & 0.375 & 3 & 0.375 & 0.5 & 3 & 0.375 & 1 & 1.5 \\ 0.375 & 0.375 & 3 & 3 & 0.666 & 3 & 0.375 & 0.666 & 1 \end{bmatrix}$$

When CCM (μ=0. 5 and α=0. 5) =D. The Eigen value and Eigen vector can be obtained from the equation (A-λ I)=0 . The maximum Eigen value is 10.23 and the corresponding Eigen vector can be normalized.

The Normalized weight of the attributes = (0.201129, 0.0276382, 0.101050 , 0.0869799, 0.034127, 0.299065, 0.0791454.0.091833)

6.2 Consistency Ratio Check

The Consistency ratio= $\frac{consistency\ index(CI)}{RI}$ where Consistency index(CI) = $\frac{\lambda_{max}-n}{n-1}$

$\lambda_{max} = 10.23$, n=9

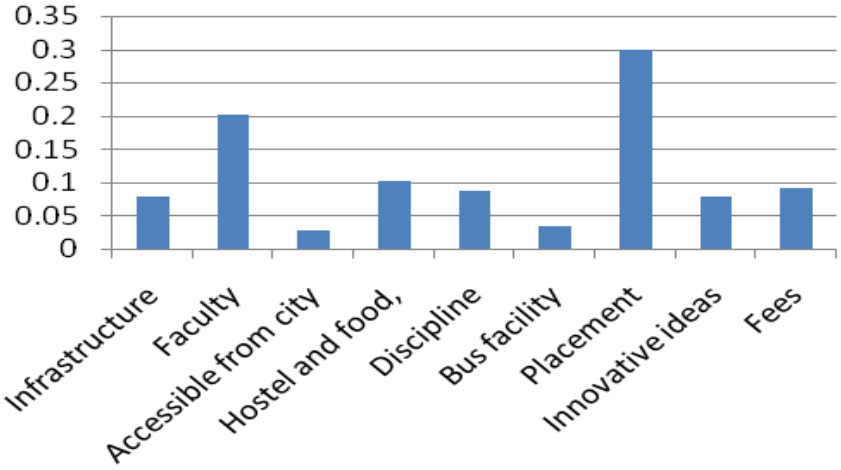
Maximum Eigen Value =10.2303 C.I.= 0.1401, C.R=0.096

So the C.R is less than 0.1 therefore the above comparison is acceptable.

Table 5

Notation	Attributes	Weights	Preference
A1	Infrastructure	0.0790305	6
A2	Faculty	0.201129	2
A3	Accessiblity from city	0.0276382	9
A4	Hostel and food, canteen	0.101051	3
A5	Discipline	0.0869799	5
A6	Bus facility	0.0341272	8
A7	Placement	0.299065	1
A8	Innovative ideas	0.0791454	7
A9	Fees	0.0918334	4

Diagram-1



7 Empirical Result

According to the literature review, and the FAHP we can get the students' expectation from the professional educational institution. From the Table-5 and diagram-1 we can get the following result .The first preference is placement , second -Faculty profile, third -Hostel and food, canteen , forth - Fees, Fifth - Discipline, sixth - Infrastructures, seventh - Innovative ideas, eighth - Bus facilities, and the last one is Distance from the city according the above table. So from the above table it clearly shows the expectations students from the professional institution using FAHP in prioritizing the student expectations.

8 Conclusion

This study investigates the key factors of student expectation in the professional education selection by applying Fuzzy membership function (Triangular) in the comparison matrix in the Analytical Hierarchy Processes. Totally nine factors are used with the survey taken from the experts in the professional institutions. Using the nine factor evaluation criteria FAHP were used to obtain the targeted solution. The fuzzy AHP approach is very useful methodology to convert user's emotions into usable design data. So according to present status the student's expectation from the professional educational Intuition the Government and private professional college have to concentrate on the Placement but once the student has mastered of the subject with soft skill the Institution can place the student very easily. But the present scenario may be will change after ten years. Using this FAHP we can obtain further more comparison we can solve many complex problems. As well we can use FAHP extension principle for further research in this same area or some other area.

References

1. Saaty, T.L.: The analytic hierarchy process. McGraw-Hill, New York (1980)
2. Triantaphyllou, E., Mann, S.H.: Using The Analytic Hierarchy Process For Decision Making In Engineering Applications: Some Challenges'. Inter'l Journal of Industrial Engineering: Applications and Practice 2(1), 35–44 (1995)
3. Kwong, C.K., Bai, H.: A fuzzy AHP approach to the determination of importance weights of customer requirements in quality function deployment (2001)
4. Yadav, H.C., Jain, R.: Prioritized aesthetic attributes of product: A fuzzy-AHP approach. International Journal of Engineering Science and Technology 4(4) (April 2012)
5. Catak, F.O., Karabas, S., Yildirim, S.: Fusy Analytic Hierarchy Based DBMS Selection In Turkish National Identity Card Management Project (2012)
6. Hsu, Y.-L., Lee, C.-H., Kreng(2010) The, V.B.: The application of Fuzzy Delphi Method and Fuzzy AHP in lubricant regenerative technology selection (2010), http://www.elsevier.com/locate/eswa
7. Saaty, T.L.: Fundamentals of Decision-Making and Priority Theory with the AHP. RWS Publications, Pittsburg (1994)
8. Kong, F., Liu, H.: Appling fuzzy analytic Hierarchy process to Evaluate Success Factors of E-Commerce. International Journal of Information and Systems Science 1(3-4), 406–412 (2005)
9. Ahmed, A., Kusumo, R., Savci, S.: Application of Analytical Hierarchy Process and Bayesian Belief Networks for Risk Analysis. Complexity International 12 (2005)
10. Zeki AYAˇ G- IIE.: A fuzzy AHP-based simulation approach to concept evaluation in a NPD environment byTransactions 37,827–842 (2005), CopyrightC IIE- ISSN: 0740-817X, print / 1545-8830 online-DOI: 10.1080/07408170590969852 (2004)
11. Hsu, Y.-L., Lee, C.-H., Kreng, V.B.: The application of fuzzy Delphi Method and Fuzzy AHP in lubricant regenerative technology selection. Expert Systems with Applications 37(1), 419–425 (2010)
12. Guo, L.: A Research on Influencing Factors of Consumer Purchasing Behaviors in Cyberspace. International Journal of Marketing Studies 3(3) (August 2011)
13. Shaverdi, M., Barzin, P.: Applying fuzzy AHP to determination of optimum selection method for economic cocoon traits improvement in silkworm breeding. Business Systems Review 1(1) (2012) ISSN: 2280-3866
14. AğIrgün, B.: Supplier Selection Based on Fuzzy Rough-AHP and VIKOR, vol. 2, pp. 1–11. Nevşehir Üniversitesi Fen Bilimleri Enstitü Dergisi (2012)
15. Wang, D., Zhang, H.: A Comparison of Fuzzy-AHP and Rough Set in Abnormal Weather Prediction. Journal of Computational Information Systems 8(14), 5991–5998 (2012)
16. Chen, Y.-C., Yu, T.-H., Tsui, P.-L., Lee, C.-S.: A fuzzy AHP approach to construct international hotel spa atmosphere evaluation model. Springer Science+Business Media Dordrecht (2012) Qual Quant doi: 10.1007/s11135-012-9792-2
17. Koul, S., Verma, R.: Dynamic Vendor Selection: A Fuzzy Ahp Approach. International Journal of the Analytic Hierarchy Process 4(2) (2012) ISSN 1936-674
18. Ishizaka, A., Nguyen, N.H.: Calibrated Fuzzy AHP for current bank account selection. Expert Systems with Applications 40(9), 3775–3783 (2013)
19. Lee, S.K.: A fuzzy analytic hierarchy process (AHP) /data envelopment analysis (DEA) hybrid model for efficiently allocating energy R&D resources: In the case of energy technologies against high oil prices, 1364-0321/ - see front matter & 2013 Elsevier Ltd. All rights reserved, http://dx.doi.org/10.1016/j.rser.2012.12.067

20. Verma., A., Gangele, A.: Investigation with Fuzzy Analytic Hierarchy Process of Green Supply Chain Management. International Journal of Innovative Technology & creative Engineering 2(5) (2012) ISSN:2045-8711
21. Chou, C.-C., Yu, K.-W.: Application of a New Hybrid Fuzzy AHP Model to the Location Choice, Mathematical Problems in Engineering Volume 2013, Article ID 592138, 12 pages. Hindawi Publishing Corporation (2013), http://dx.doi.org/10.1155/2013/592138
22. Rouhani, S., Ashrafi, A., Afshari, S.: Segmenting Critical Success Factors for ERP Implementation Using an Integrated Fuzzy AHP and FuzzyDEMATEL Approach. World Applied Sciences Journal 22(8), 1066–1079 (2013) ISSN 1818-4952, doi: 10.5829/idosi.wasj.2013.22.08.631
23. Aggarwal, R., Singh, S.: AHP and Extent Fuzzy AHP Approach for Prioritization of Performance Measurement Attributes. World Academy of Science, Engineering and Technology 73 (2013)
24. Tas, A.: A Fuzzy AHP approach for selecting a global supplier in pharmaceutical industry. African Journal of Business Management 6(14), 5073–5084 (2012), doi: 0.5897/AJBM11.2939, ISSN 1993-8233, http://www.academicjournals.org/AJBM
25. Astuti, R.: Risks and Risks Mitigations in the Supply Chain of Mangosteen: A Case Study. Operations And Supply Chain Management 6(1), 11–25 (2013) ISSN 1979-3561| E, ISSN 1979-3871

Qualitative Learning Outcome through Computer Assisted Instructions

Tamali Bhattacharyya[1], Rajendra Prasath[2], and Bani Bhattacharya[1]

[1] Center for Educational Technology
[2] Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India
{tamali95,drrprasath}@gmail.com, banib@cet.iitkgp.ernet.in

Abstract. The Qualitative Outcome of Learning (QOL) measures the level of attainment in learning efforts through *Structure of the Observed Learning Outcome* (SOLO) taxonomy. Modern Computer Assisted Instructions (CAI) employ interactive tools that illustrate a concept using not only visual aids like animation, sound, etc but also stimulating the cognitive level of learners. In this work, we attempted to evaluate the QOL of learners at the age of 14 years. The learners are challenged to test their understanding in lessons related to specific topics, rather than just knowing the content. Each topic content is based on SOLO taxonomy and set to test the understanding level of learners with increasing complexity in that topic. Lessons are planned from simple to complex structured responses involving Relational and abstract thinking and the level at which the learner is operating is assessed. Our objective is to evaluate the QOL using CAI tools whether they can uplift "surface" learners towards "deep" learning. We used *Shikshak* - an intelligent tutoring system and a simple CD based approach, through which learners appear to promote active cognitive learning and assessed the growth or decline of competence of learners with the help of CAI according to SOLO taxonomy. The effects of QOL by the CAI tools have been measured in comparisons with printed lessons based learning approach and found that CAI stimulates "deep" learning.

Keywords: Qualitative Outcome of Learning, Computer Aided Instructions, SOLO Taxonomy, Deep Learning, Shikshak - Intelligent Tutoring System.

1 Introduction

Assessment is an essential part of the "teaching-learning process". A large number of studies have proved how the assessment procedures and modalities affect the level of learning outcomes reached by learners. In most systems of education, evaluation is done "Quantitatively", in terms of how much a learner has learned. However, such quantitative evaluation fails to take into account partial understanding or the level of understanding by a learner and therefore, cannot gauge the depth of learners' understanding of a subject matter. The idea of looking at learning from a qualitative point of view implies a global outlook towards learning. "In the qualitative outlook it is assumed that learners learn cumulatively and constructively, interpreting and combining new material with what they already know, thereby modifying their understanding

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 567–578, 2013.
© Springer International Publishing Switzerland 2013

progressively as they learn"[4,3,1,5,6,2,9] Research studies have shown that the SOLO taxonomy [5] could serve as a useful tool for categorizing learners' responses in a qualitative way.

Learning approaches are different and all learners do not respond equally to the same teaching technique. Learning technology is used for the enhancement of teaching, learning and assessment. It includes computer-based learning and multimedia materials to support learning. Learning technology can offer learners control over their learning and flexibility, so that one can learn in the style most effective for each individual. The ideal package will offer a highly structured route of study for different levels of learners with different learning approaches - deep, achieving and surface. The focus of this study is on the change in the qualitative learning outcome after using the three different learning tools - intelligent tutoring system, CD based learning tool and Printed Text material using the "Structure of the Observed Learning Outcome" (SOLO) taxonomy.

2 Objective

This study aims at investigating the role of instructional design of the contents using SOLO taxonomy, different types of learning tools and study approaches on the level of understanding (Qualitative learning) in Geometry in Central Board of Secondary Education (CBSE) affiliated learners of class 9 learners in India. Here the objective is to assess the qualitative learning outcome of a learner in the secondary school level with the help of computer aided instructions. Our study aims to capture information on learners from the secondary schools to investigate change in their qualitative learning outcomes with the help of computer aided learning, that is, intelligent tutoring system and CD based approach. Simultaneously the investigation is conducted through printed textual material to determine the change, if any, in the qualitative learning outcomes. The focus of this research would be whether the software can uplift surface learners and orient them towards deep learning.

3 Methodology

3.1 Participants

To determine the QOL of the learners, three experiments were envisaged and arranged. In the first experiment, an exclusive workshop was organized in order to evaluate the change in the QOL and learning approaches of the learner after using an "Intelligent Tutoring System" (ITS), called Shikshak. Thirty-three learners from class IX under the Central Board of Secondary Education (CBSE) participated in this experiment. Of these learners, eleven were girls and twenty two were boys. This workshop was conducted for three weeks and the learners spent an hour a day on two Geometry topics related to line-plane and Triangle. These thirty three learners were already taught these topics by their respective subject teachers and were quite comfortable with the computer infrastructure provided during the workshop because they had similar facilities in their schools. The learning materials in the repository were annotated with a set of meta-data which helped in retrieving the appropriate learning material according to the requirement

of the learners with different learning approaches. The ITS Shikshak supports various kinds of materials like text, audio, video, presentations, etc. in any standard format (e.g. Microsoft Word or Microsoft PowerPoint documents, PDF documents, Macromedia Flash movies, etc.).

The second experiment was organized among learners of Standard IX, in a CBSE affiliated school (Kendriya Vidyalaya, Salua), spread over two sections, where the learners used the same contents in CDs. The same Geometry contents were structured according to the pedagogic learning principles. Here fifty two learners participated; out of which thirty three were girls and remaining twenty nine were boys. This workshop was conducted for three weeks and the learners spent one hour each day.

The third experiment was conducted in the same school where twenty five learners from a third section participated, out of which eleven were girls and fourteen were boys. They were provided with the same contents, but in the printed textual form.

3.2 Instruments

Four instruments were used in order to carry out the study. The first instrument is to structure the contents according to the SOLO taxonomy. The SOLO taxonomy has been found to be an ideal tool for qualitatively assessing learners in higher education. It was developed by Biggs and Collis (1982) and is well described in Biggs (1999)[5]. The taxonomy makes it possible, in the course of learning, teaching, or assessing a subject, to identify in broad terms the level at which a learner is currently operating. The taxonomy consists of five categories of increasingly complex stages: The categories are: Prestructural; Unistructural; Multistructural; Relational and Extended Abstract. For the purpose of this study, we have not taken into consideration the first elementary stage of "Prestructural level". The most desirable level is the Extended Abstract level. Metacognition enables us to be successful learners who could possibly reach the level of the Extended Abstract and has been associated with intelligence (e.g., Borkowski et al. , 1987) [7]. It refers to higher order thinking which involves active control over the cognitive processes engaged in learning. It simply defined as "thinking about thinking."

The second instrument used is the Computer Aided Instruction (CAI). It refers to instruction or remediation presented on a computer, employing interactive tools and illustrates a concept through attractive animation, sound and demonstration. Experts in education are actively involved in developing ways for learners to use technology to improve education. This experiment envisages examining the QOL of the learner with the help of CAI tools to explain how they can assist in implementing strategies and learning activities supported by the pedagogical approaches. To support the CAI, we used two tools - intelligent tutoring system (ITS) and simple CD based system where the contents are structured according to the pedagogic learning principles and illustrate a concept through attractive animation, sound and demonstration.

The third instrument is the printed text document in which the Geometry contents are structured according to the instructional design based on SOLO taxonomy. The text materials comprised the same contents in the form of eighty nine text documents and printed version of twelve Power Point presentations and twenty two flash movies for this entire Geometry course.

The fourth instrument is the achievement test. The achievement instruments used were the pre-test and post-test. The same test instruments were used for the pre-test and post test, however the post - test questions were arranged in the same order as they were presented on the pre-test. Eight evaluation tests were conducted, out of which four for pre-tests and four for post tests. Each test contained the same content questions on Geometry - Line plane and Triangle. The test instrument was objective and was designed using different questioning formats, in order to appeal to a variety of learner's testing strengths. The instrument was given to each learner in the sample and took approximately thirty minutes to complete. The evaluation questions are designed from lower order to higher order thinking-based on SOLO taxonomy. The taxonomy of competence, from low order to higher order, includes Unistructural, Multistructural, Relational and Extended Abstract level. These competencies are also beneficial for creating test questions.

3.3 Research Model

The focus of the study is to investigate primarily on learning from CD based technology with multimedia instructional methods and Shikshak, an intelligent tutoring system, with which learners may appear to promote active cognitive learning. The contents designed in Geometry for both CD based technology and intelligent tutoring system. The purpose is to assess the growth or decline of competence of learners with the help of CAI according to the taxonomy. In the study, a new item format was designed and a new criterion framework of assessment based on Biggs' SOLO Taxonomy. Each topic is designed with the help of the SOLO taxonomy which describes level of increasing complexity in a learner's understanding of the subject. The same content provided in a printed textual form so that one can make a comparative study to identify the learning technology which promotes deep learner learning.

In this model, the Dependent Variable consists of the level of "Qualitative level of Learning Outcome" (QOL) in Geometry as determined by tests devised on the basis of the 'Structure of Observed Learning Outcome' (SOLO) Taxonomy.

Dependent Variables:
Qualitative Outcome of Learning (SOLO Categories):

1. Unistructural level
2. Multistructural level
3. Relational level
4. Extended Abstract level.

Independent Variables:

1. Learning Strategy (Different types of learning tools)
2. Instructional design of the content

QOL = f(Learning Strategy using different learning tools, instructional design of the content)

As it is an experimental study and the purpose is to see the effectiveness of independent variables, that is, learning strategy and instructional design of the content to the effect on QOL. We use QOL as an outcome measure to exemplify an evaluation of two delivery modes: traditional learning tools and computer aided instructions. To compare two different types of delivery modes, we begin with a simple framework linking the pedagogical environment to learning outcomes. The learner learning outcome is the product of a cumulative process that takes place over a period of time - for example between the beginning and end of the experiment. Using test scores as a proxy for learning, we can express change in the qualitative level of learning outcome

4 Learning Tools

The Shikshak - An Intelligent Tutoring System (ITS)[1] is a computer-based tutor which acts as a supplement to human teachers. The software tracks learners' work, tailoring feedback and hints along the way. An ITS called *Shikshak* (Chakraborty et al., 2007)[8] was built, where there are different modules performing different tasks of teaching. There are three different modules in Shikshak - domain model, learner model and teaching model. Domain model contains a repository for study and test materials. It is also compatible with various kinds of documents, such as, explanation type, experiment type, application type, exercise type, etc. (Roy *e*t al, 2007)[11] and also of various formats (text, audio, video, flash movies, presentation, etc.). Hence, the repository is capable of storing different documents and we stored different documents of class IX Geometry, for the topics related to line, plane and angle according to the learning approaches. The teaching method planned by the system is done on the basis of the data available in the learner model. Pedagogical model is the nucleus of the system which controls the actual teaching process. The major tasks of this module include customized material selection from the repository, analyzing the feedbacks (from the test results) from the learners and updating the learner model and also making necessary changes in the teaching plans to suit the learner better in the subsequent lessons. For example, from the learner model, if a learner's learning approach is found to be deep, surface or achieving level, the pedagogical model will try to select those materials which are suited for his/her learning approach. The overall architecture is given in figure 1.

Multimedia Content CD is very simple and easy to use. Here a simple HTML structure has been used and a content tree structure has been developed. Each topic has four levels - Unistructural, Multistructural, Relational and Extended Abstract according to the SOLO Taxonomy. For each topic, the learner starts the course at the Unistructural level, where he/she is presented a set of contents. After the learner has gone through the set, he/she has to answer an interactive review test, passing which he/she goes one level above that is the Multistructural. The same process is repeated for every level.

The Printed Text material also follows the same format as that of the CD based system. learners are given Printed Text material for each level in every topic. At the end of each level, they answer the multiple choice questions to graduate to the next level.

[1] *Shikshak* and *ITS* are interchangeably used in this paper.

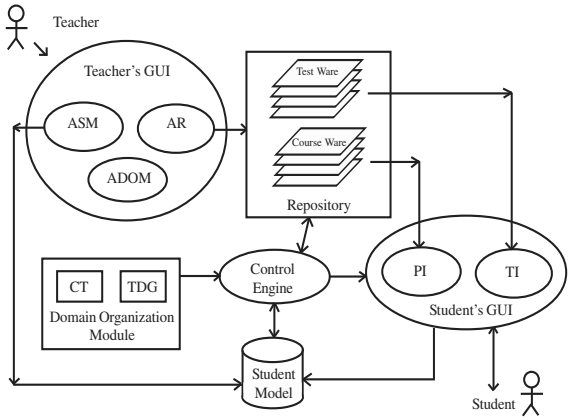


Fig. 1. Overall System Architecture of Shikshak - reproduced from [8]; ASM: Authoring learner Model; AR: Authoring Repository; ADOM: Authoring Domain Organization Module; CT: Content Tree; TDG: Topic Dependency Graph

4.1 Course Format Description

Problem-based lessons promote active learner learning and deep thinking of mathematics. We decided to focus our concept in Geometry for spatial reasoning and problem solving. Integrating pedagogy and subject knowledge through experiencing a variety of tasks for learners, including interactive activities on computer aided learning, leads to successfully engaging all learners in geometric thinking. learners can take advantage of technology for experimenting, conjecturing, and exploring topics in depth.

We focused on two basic lessons in Geometry - Line plane and Triangle. The Line plane has four parts: Straight line, Angle, Parallel lines and Sum angles. The Triangle has three parts: - Congruency of a Triangle, Isosceles, Inequality of a Triangle. The design of each topic is based on SOLO taxonomy. SOLO describes level of increasing complexity in a learner's understanding of a subject through four stages and it claims to be applicable to any subject area. According to this taxonomy, each part of these two lessons line-plane and Triangle were structured into four parts. Those parts are Unistructural, Multistructural, Relational and Extended Abstract. Each part was presented in simple textual form and also in different multimedia presentations like Power Point presentations and animated presentations. All these different kinds of documents were gathered from various sources, such as several web sites, class notes from experienced subject teachers, educational CDs and course material developed with the help of different books.

Four experienced mathematics teachers were chosen for the evaluation of the contents. The first step was to explain the guidelines of Bloom and SOLO taxonomy. The second step constituted in the evaluation of these contents by these teachers to check whether the contents conformed to the respective taxonomies. There were 89 text documents, twelve Power Point presentations and twenty two Flash movies for this entire Geometry course.

In each module, the learner had to go through summative and formative evaluation process. Here in each module, the learner has to go through multiple choice questions

and the purpose is to identify aspects of performance that need to improve and to offer corrective suggestions. But to find out the QOL, we mainly focused on summative evaluation. It is a process of identifying larger patterns and trends in performance and judging these summary statements against criteria to obtain performance ratings.

4.2 Objectives

a) Is there a significant difference in the level transitions among learners using ITS, CD based and printed text materials?
b) Is there a significant difference in the QOL using the intelligent tutoring system?
c) Is there a significant difference in the QOL using CD based animation graphic courseware to the learners?
d) Is there a significant difference in the QOL among learners using Printed Text material following SOLO taxonomy?

5 Empirical Results

After using the ITS, CD based system and Printed Text material, it was observed that transition of learners from one learning outcome to the other can be described in the following pattern as illustrated in Figure 2.

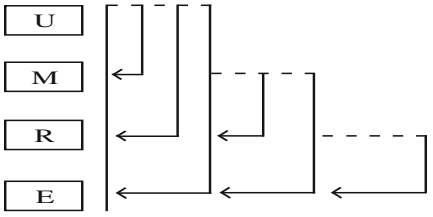


Fig. 2. Transition of learners from One Level to Another

Three types of transitions can occur from Unistructural(U) to either Multistructural(M) or Relational(R) or Extended Abstract(E):

Unistructural to Multistructural(UM)
Unistructural to Relational(UR)
Unistructural to Extended Abstract(UE)

Two types of transitions can occur from Multistructural(M) to either Relational(R) or Extended Abstract(E):

Multistructural to Relational (MR)
Multistructural to Extended Abstract(ME)

One type of transition can occur from Relational(R) to Extended Abstract(E):

Relational to Extended Abstract(RE)

Another possibility is that the learner can remain in their respective learning outcome levels, in which case the resultant combinations would be:

Unistructural to Unistructural(UU)
Multistructural to Multistructural(MM)
Relational to Relational(RR)
Extended Abstract to Extended Abstract(EE)

The change in the transition in the learning process can be seen in the Table. 1. This matrix can be decomposed into lower and upper diagonal matrix along the left diagonal. We hardly focus on the lower diagonal. The diagonal and upper diagonal describe the transitions levels in the learning improvement across the lower levels to higher levels.

Table 1. Transition Levels in the Learning Process

UU	**UM**	**UR**	**UE**
MU	**MM**	**MR**	**ME**
RU	RM	**RR**	**RE**
EU	EM	ER	**EE**

During the evaluation, we have used Stat-Ease statistical software - *Design-Expert* to analyze Randomized Two Level Factorial Design with no blocks and measured two factor interaction effects. Theoretical details behind the two level factorial design can be read from [10]. The first factor includes 7 types of topics on Geometry: *straight line, angle, parallel lines, sum of angles, congruency, isosceles, inequality* and the second factor includes 4 levels of qualitative learning outcome (Unistructural (U), Multistructural(M), Relational (R) and Extended Abstract(E)) according to SOLO taxonomy. These 4 levels are considered as baseline levels before our experiments. Here we have observed that among 110 learners, 33 learners used ITS, 52 learners used CD based approach and 25 learners used printed TEXT based approach.

We measured the qualitative outcome of learning after using three different learning tools and this measurement is guided by the following observations:

i) The change in the level of the learners in U is measured by the changes in $U \rightarrow U = UU$.
ii) The change in the level of the learners in M is measured by the changes in $U \rightarrow M\ AND\ M \rightarrow M = UM + MM$.
iii) The change in the level of the learners in R is measured by the changes in $U \rightarrow R\ AND\ M \rightarrow R\ AND\ R \rightarrow R = UR + MR + RR$.
iv) The change in the level of the learners in E is measured by the changes in $U \rightarrow E\ AND\ M \rightarrow E\ AND\ R \rightarrow E\ AND\ E \rightarrow E = UE + ME + RE + EE$.

The overall objectives of the underlying experimental design and evaluation include: 1) The learning outcome is measured basically by counting the improvement obtained over the qualitative learning outcome from lower to higher level with its significant improvements which is obtained by two level factorial design with no blocks and 2) Outlier detection through the calculation of residuals and the normal probability plot.

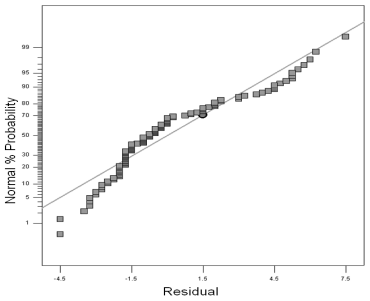


Fig. 3. NP Plot before using learning tools

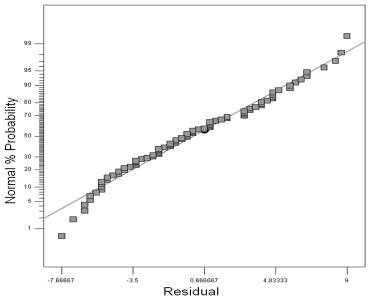


Fig. 4. NP Plot after using learning tools

At first, we have carried out an experiment with a full factorial design with no blocks and 84 observations. The design model used is 2 level factorial design and we have computed the Analysis of Variance(ANOVA) with main effects (7 topics in Geometry and 4 levels) and interaction effects (mixture of 4 levels according to SOLO taxonomy). The experiments were done to measure the QOL before and after the use of 3 different tools - Shikshak(ITS) based learning, CD based learning and printed TEXT based learning. Now we compare the transition of learning of learners before and after 3 different learning tools.

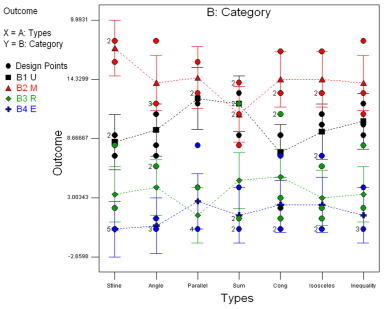


Fig. 5. IG before using 3 Learning Tools

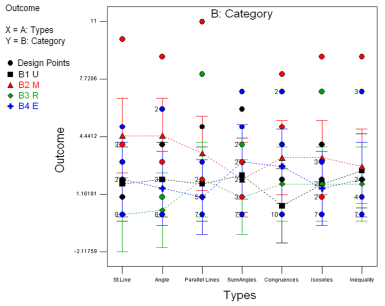


Fig. 6. IG after using 3 Learning Tools

Figure. 3 shows the Normal Probability (NP) plot of the learning outcome before applying 3 learning tools. In this experiment, we have observed that most of the learners were found in U and M categories. From the computed ANOVA, the $F-$value of 8.00 implies that the model is significant. Now we validate our hypothesis of measuring the qualitative learning outcome after applying 3 different learning tools whose content follows instructional design with SOLO taxonomy. The same experimental setup is followed and ANOVA table is computed after applying the learning tools. The model $F-$value of 1.83 implies that the model is significant. Also the values of $Prob > F$ shows that the there are significant improvements in the levels of learning at 95% level of significance. This reflects in Figure. 4 where outliers are much more reduced and most of the learners lie on the best fit line. Additionally, the interaction effects within and across each type of topics are illustrated through interaction graphs. The actual

improvements in the learning process are shown by two design points for each levels of learning outcome and '+' mark indicates the actual improvement in the level of learning outcome. This means that the pedagogic instructional design (SOLO taxonomy as given in 3.3) makes instructional content more engaging and efficient which, in turn, improves QOL. The corresponding interaction graphs(IG) are given in Figure. 5 and Figure 6.

Table 2. The QOL before ITS (according to SOLO taxonomy)

Term	DF	Sum of Squares	Mean Square	%Contribution
Types (7 Topics)	6	13.71	2.29	2.56
Levels(4 Levels)	3	323.25	107.75	60.42
Interaction	18	198.00	11.00	37.01

Next we have measured the qualitative outcome of learning using each individual tool. First we consider ITS based approach and its overall improvement in the learning outcome. Before using the ITS system, the overall contribution of QOL according to the SOLO taxonomy is 60.42% as given in the Table. 2. Figure. 7 refers to the corresponding Normal Probability plot illustrating the effects of the qualitative learning outcome before ITS.

After using ITS based learning approach, the overall contribution of QOL according to the SOLO taxonomy reaches to 86.94% as given in the Table. 3. This improvement is due to the outcome of the experiment with the factorial design chosen with interaction effects which are highlighted in the transition level matrix 1 (given in bold). This is also illustrated by the corresponding normal probability plot in Figure. 7, 8, 9 and 10.

Table 3. The QOL after ITS (according to SOLO taxonomy)

Term	DF	Sum of Squares	Mean Square	%Contribution
Types (7 Topics)	6	0.0	0.0	0.0
Levels(4 Levels)	3	560.96	186.99	86.94
Interaction	18	84.29	4.68	13.06

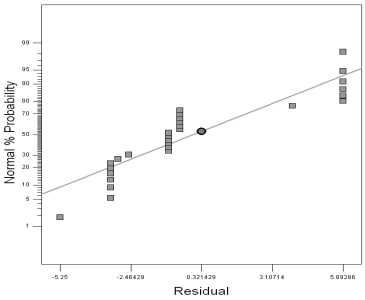


Fig. 7. Normal Probability Plot of Before ITS based learning

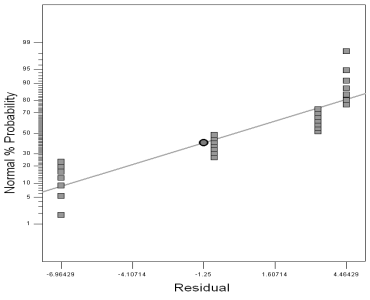


Fig. 8. Normal Probability Plot of After ITS based learning

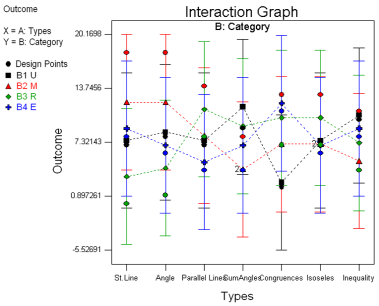


Fig. 9. IG before ITS based learning

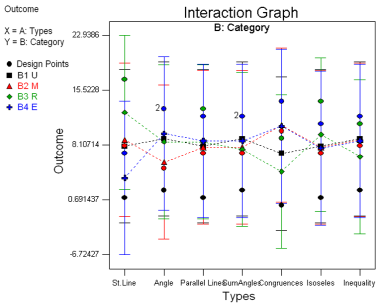


Fig. 10. IG after ITS based learning

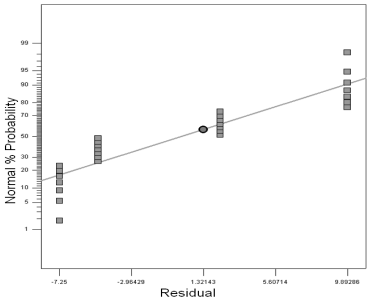


Fig. 11. Normal Probability Plot of Before CD based learning

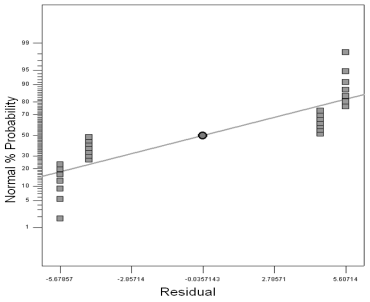


Fig. 12. Normal Probability Plot of After CD based learning

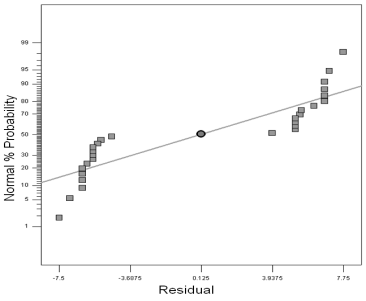


Fig. 13. Normal Probability Plot of Before printed TEXT based learning

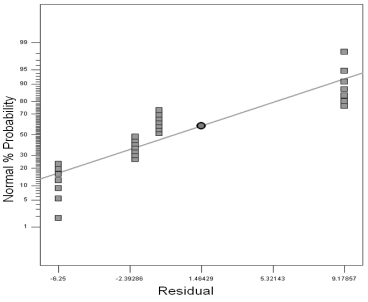


Fig. 14. Normal Probability Plot of After printed TEXT based learning

Before CD based approach, the $F-$value of QOL obtained through the interaction among the factors as 0.063(refer to Figure. 11). In CD based learning approach, we have noticed a shift in the learning outcome from M to R level (which refers to combined interactions effects of MR and RR). The $F-$value of 0.21 shows that there is 3 fold improvements in the qualitative outcome of learning. This is clearly shown in the Normal probability plot as given in Figure. 12.

Before printed TEXT based approach, even though the improvement is not significant in the learning outcome of individual levels, the interaction effects improved. This is clearly shown in Figure. 13 and Figure. 14.

6 Conclusion

In this paper, we have attempted to evaluate the Qualitative Outcome in the Learning process of learners at the age of 14 years. The learners are challenged to test their deeper level of understanding in lessons related to specific topics. Each topic content is based on SOLO taxonomy and lessons are planned from simple to complex structured responses involving Relational and abstract thinking. We conducted the experiments with the content designed using the instructional design (SOLO taxonomy). We observed from the factorial design experiment with the focus on interaction effects that the qualitative outcome of learning improved over subsequent levels with the help of three different learning tools and learners attained significant improvements over the learning process towards deep learning through each of the tools.

References

1. Biggs, J.: Assessing for learning: Some dimensions underlying new approaches to educational assessment. The Alberta Journal of Educational Research 41, 1–17 (1995)
2. Biggs, J.: Enhancing teaching through constructive alignment. Higher Education 32, 347–364 (1996), http://dx.doi.org/10.1007/BF00138871
3. Biggs, J.B.: Australian Council for Educational Research Limited. learning Process Questionnaire MANUAL - Student Approaches to learning and studying (1987)
4. Biggs, J.B.: The Process of Learning, 2nd edn. Prentice Hall, Sydney (1987)
5. Biggs, J.B.: Teaching for Quality Learning at University: What the student does. Open University Press, Buckingham (1999)
6. Biggs, J.B., Collis, K.: Evaluating the quality of learning: The S. O.L.O. Taxonomy. Academic Press, New York (1982)
7. Borkowski, J.G., Carr, M., Pressley, M.: "spontaneous" strategy use: Perspectives from metacognitive theory. Intelligence 11(1), 61–75 (1987)
8. Chakraborty, S., Bhattacharya, T., Bhowmick, P., Basu, A., Sarkar, S.: Shikshak: An intelligent tutoring system authoring tool for rural education. In: International Conference on Information and Communication Technologies and Development, ICTD 2007, pp. 1–10 (December 2007)
9. Marton, F., Saljo, R.: On qualitative differences in learning - 1: Outcome and process. British Journal of Educational Psychology 46, 4–11 (1976)
10. Montgomery, D.C.: Design and Analysis of Experiments. John Wiley & Sons (2006)
11. Roy, D., Sarkar, S., Ghose, S.: Learning material annotation for flexible tutoring system. Journal of Intelligent Systems 16, 293–305 (2007)

Bayesian Classification for Image Retrieval Using Visual Dictionary

M. K. Nazirabegum[1] and N. Radha[2]

[1] PSGR Krishnammal College for women, Coimbatore
nazirabegum17@gmail.com
[2] GRG School of Applied Computer Technology, Coimbatore

Abstract. Image Retrieval is one of the most promising technologies for retrieving images through the query image. It enables the user to search for the images based upon the relevance of the query image. The main objective of this paper is to develop a faster and more accurate image retrieval system for a dynamic environment such as World Wide Web (WWW). The image retrieval is done by considering color, texture, and edge features. The bag-of-words model can be applied to image classification, by treating image features as words. The goal is to improve the retrieval speed and accuracy of the image retrieval systems which can be achieved through extracting visual features. The global color space model and dense SIFT feature extraction technique have been used to generate a visual dictionary using Bayesian algorithm. The images are transformed into set of features. These features are used as an input in Bayesian algorithm for generating the code word to form a visual dictionary. These code words are used to represent images semantically to form visual labels using Bag-of-Features (BoF). Then it can be extended by combining more features and their combinations. The color and bitmap method involves extracting only the local and global features such as mean and standard deviation. But in this classification technique, color, texture, and edge features are extracted and then Bayesian Algorithm is applied on these image features which gives acceptable classification in order to increases the accuracy of image retrieval.

Keywords: Image retrieval, Visual dictionary, Genetic algorithm, Bayesian algorithm.

1 Introduction

The earlier image retrieval systems were focused on the text that is annotated for a particular image. It is not efficient to search an image based upon the text in the order of it, as all the images are not completely tagged. Without the ability to examine image content, explore must rely on metadata such as captions and keywords, which may be difficult or costly to produce. To search for a relevant image, the exact content properties of the image are needed. The properties are the visual features such as color, shape, texture, regions etc. These features provide valuable and high percentage for mapping techniques.

Section 2 describes the existing works of image retrieval systems. Section 3 describes about the techniques used for retrieving image from the database. Section 4

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 579–587, 2013.
© Springer International Publishing Switzerland 2013

describes methodology for image retrieval. Section 5 describes the results. Section 6 provides conclusion and Section 7 provides future enhancement.

2 Literature Survey

2.1 Texture Based Image Indexing and Retrieval

N. Rao, et al. [3] proposed the system as there is a prominent increment in computing power, rapidly reducing storage cost and worldwide access to the Internet. Digital information has become popular in the recent years. Digital information is preferable to analog formats because of convenient sharing and distribution properties. This trend has motivated research in image databases, which were nearly ignored by traditional computer systems due to the enormous amount of data necessary to represent images and the difficulty of automatically analyzing images. Currently, storage is less issue because huge storage capacity is available at low cost. However, effective indexing and searching of large-scale image databases remains as a challenge for computer systems. The image retrieval is a system, which retrieves the images from an image collection where the retrieval is based on a query, which is specified by content and not by index or address. The query image is an image in which a user is interested and wants to find similar images from the image collection.

2.2 Content-Based Image Indexing and Searching

J. Z. Wang and G. Wiederhold proposed a method for content based image indexing [4]. Every day, large numbers of people are using the Internet for searching through different multimedia databases. In current real-world image databases, the prevalent retrieval techniques involve human-supplied text annotations to describe image semantics. These text annotations are then used as the basis for searching, using mature text search algorithms that are available as free-ware. However, there are many problems in using this approach. For example, different people may supply different textual annotations for the same image. This makes it extremely difficult to reliably answer user queries. Furthermore, entering textual annotations manually is excessively expensive for large-scale image databases.

2.3 Color Image Retrieval Technique Based on Color Features and Image Bitmap

T. C. Lu and C. C. Chang [5] proposed an image retrieval technique based on color features and image bitmap. The field of color image retrieval has been an important research area for several decades. The global characteristics of the image such as color distributions, the mean value and the standard deviation are used to retrieve more similar images effectively from the digital image databases. Moreover, the image bitmap is used to represent the local characteristics of the image for increasing the accuracy of the retrieval system. In a CBIR system, images are automatically indexed by summarizing their visual contents through automatically extracted primitive features such as shape, texture, color, size, and so on.

Each image in the image database may be different from all the others, but at the same time all images may share certain common characteristics. Hence, there is a need for the statistical description of images to capture these common characteristics and use them to represent an image with fewer bits. The statistical description has been developed using mean values and standard deviations of images [5]. In the proposed scheme, each pixel of a color image is represented by a vector

$$P_i = \begin{bmatrix} R_i \\ G_i \\ B_i \end{bmatrix} \tag{1}$$

2.4 Genetic Algorithm Based Image Retrieval

S. F. da Silva, M. A. Batista, and C. A. Z. Barcelos, [8] proposed an adaptive image retrieval through the use of a genetic algorithm. A Genetic algorithm works with a population of individuals. It represents the possible solutions to a given problem. The population of individuals is usually randomly generated. However, if there is a knowledge available concerning the problem domain (heuristic), it can be incorporated into a fraction of the initial set of potential solutions. The evolution process halts when the system no longer improves, or when a preset maximum number of generations are reached. The output of the genetic algorithm is usually the best individual of the end population. Digital image libraries and other multimedia databases have been dramatically expanded in the recent years. In order to effectively and precisely retrieve the desired images from a large image database, the development of a content-based image retrieval (CBIR) system has become an important research issue. In this document, a user-oriented mechanism for CBIR method based on an interactive genetic algorithm (IGA) is proposed. Color attributes like the mean value, the standard deviation, and the image bitmap of a color image are used as the features for retrieval. In addition, the entropy based on the gray level co-occurrence matrix and the edge histogram of an image is also considered as the texture features. Furthermore, to reduce the gap between the retrieval results and the users' expectation, the IGA is employed to help the users to identify the images that are most satisfied to the users' need.

3 Image Retrieval

The recent emergence of multimedia databases and digital libraries makes retrieval of image data based on pictorial representations more ease. While manual image annotations can be used to a certain extent to help image search, the feasibility of such an approach to large databases is a questionable issue. Simple textual descriptions can be ambiguous and often inadequate for database search. There are two general types of image search namely target search and category search. The goal of the target search is to find a specific (target) image, such as a registered logo, historical photograph, or a particular painting. The goal of category search is to retrieve a given semantic class or genre of images, such as scenery images. In other words, a user uses target search to find the known image. In contrast, category search is used to find relevant images. This document mainly focuses on category search technique.

3.1 Image Retrieval System Based on Interactive Bayesian Algorithm

In this scheme, the visual feature extraction process for retrieving similar images is proposed. The features extracted are stored in a .cfs file. Various features are considered that includes color, shape, texture etc. These features act as a key term for matching the related images. The color features are matched based upon the intensity, hue and saturation values are calculated based upon the colors can be recognized. The shape is recognized where the image is divided in to several regions. Then all the pixels on the region boundaries are found. It includes various steps like smoothening, edge detection, removing false edges(noise) then sharpening the edge, filling up the gaps and then putting boundary pixels in order.

The color correlogram are characterized not only the color distributions of pixels, and also the spatial correlation of pairs of colors. The first and second dimensions of the three-dimensional histogram are the colors of any pixel pair and the third dimension is their spatial distance. A color correlogram is a table indexed by color pairs, where the k^{th} entry for (c_i, c_j) specifies the probability of finding a pixel of color c_j at a distance k from a pixel of color c_i in the image.

Color Descriptor Based Image Retrieval

A color image can be represented using three primaries of a color space. Since the RGB space does not correspond to the human way of perceiving the colors and does not separate the luminance component from the chrominance ones, In this approach the HSV color space model is used. HSV is an intuitive color space in the sense that each component contributes directly to visual perception, and it is common for image retrieval systems Hue is used to distinguish colors, whereas saturation gives a measure of the percentage of white light added to a pure color.

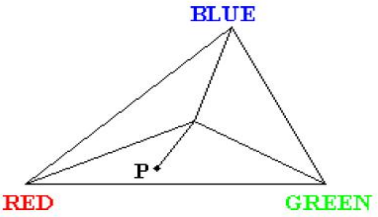


Fig. 1. Obtainable HSV color from RGB color space

The above Fig. 1 shows obtainable HSV color from RGB color space. In the obtainable HSV colors lie within a triangle whose vertices are defined by the three primary colors in RGB space. The hue of the point P is the measured angle between the line connecting P to the triangle center and line connecting RED point to the triangle center. The saturation of the point P is the distance between P and triangle center. The value (intensity) of the point P is represented as height on a line perpendicular to the triangle and passing through its center. The grayscale points are situated onto the same line.

Texture Based Image Retrieval

Texture is an important attribute that refers to innate surface properties of an object and their relationship to the surrounding environment. CBIR is the best appropriate

texture descriptor. A gray level co-occurrence matrix (GLCM) [10], which is a simple and effective method for representing texture has been used.

Edge Histogram Based Image Retrieval

Edges in images constitute an important feature to represent their content. Human eyes are sensitive to edge features for image perception. One way of representing such an important edge feature is to use a histogram. An edge histogram in the image space represents the frequency and the directionality of the brightness changes in the image. The edge histogram descriptor (EHD) [11] describes an edge distribution with a histogram based on local edge distribution in an image.

The extraction process of EHD consists of the following stages.

1) An image is divided into 4 × 4 sub-images.
2) Each sub-image is further partitioned into non-overlapping image blocks with a small size.
3) The edges in each image block are categorized into five types: vertical, horizontal, 45◦ diagonal, 135◦ diagonal and non-directional edges.
4) Thus the histogram for each sub-image represents the relative frequency of occurrence of the five types of edges in the corresponding sub-image.
5) After examining all image blocks in the sub-images, the five-bin values are normalized by the total number of blocks in the sub-image. Finally, the normalized bin values are quantized for the binary representation. These normalized and quantized bins constitute the EHD.

Interactive Bayesian Algorithm

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. Bayesian classifier is a statistical classifier that performs probabilistic prediction i.e. predicts class membership probabilities. The foundation is based on Bayes theorem.

Bayesian classification is used to find out the probability of the image features. Usually features are extracted from the images. The features are namely color, texture, shape etc based upon the similarity of features the images are retrieved from the large image database. The following are the process steps involved in Bayesian Classification.

1) Querying: The user provides a sample image as the query for the system.
2) Similarity computation: The system computes the similarity between the query image and the database images according to the aforementioned low-level visual features.
3) Retrieval: The system retrieves and presents a sequence of images ranked in decreasing order of similarity. As a result, the user is able to find the relevant images by getting the top-ranked images first.
4) Incremental search: After obtaining some relevant images, the system provides an interactive mechanism via IGA(Interactive genetic algorithm), which lets the user, evaluates the retrieved images as more or less relevant to the query and the system then updates the relevant information to include as many user-desired images as possible in the next retrieval result. The search process is repeated until the user is satisfied with the result or results cannot be further improved.

Bayesian have the following advantages over traditional search methods

1) They directly work with a coding of the parameter set.
2) The search process is carried out from a population of potential solutions.
3) Payoff information is used instead of derivatives or auxiliary knowledge.
4) Probabilistic transition rules are used instead of deterministic ones.

4 The Proposed Methodology

4.1 Querying

The user provides a sample image as the query for the system. The users of the system can give an input image of a specific kind and can view the result based upon the efficient retrieval. The user must provide images as the input for the system. The user should be able to search and retrieve the related images effectively from it. By exactly learning the similarities of images, users can have a comfortable image adaptation capabilities for the smart devices which reduce the dissimilarity and less adaptable capabilities.

4.2 Bag-of-Words

In the Bag of visual model, each image is represented as a set of order less visual words. The local descriptors are much more precise and discriminating than global descriptors. When looking for a specific image in a image database or target object within the image, this discrimination is very essential in retrieval system, but when looking for complex categories in a large image database, it becomes difficult.

To minimize this problem, there is scope of improvement to propose a possible solution is the technique of visual dictionaries. It is made up of set of bag of visual words constructed using clustering approaches. In the visual dictionary label representation, each region of an image becomes a visual “word” of the dictionary. The idea is that different regions of description space will become associated with different semantic concepts. For example, sky, clouds, land, rocks, vegetation etc having its own semantics.

4.3 Quantization with Dense Sift Feature Based Retrieval

The Dense SIFT feature extraction algorithm is used to extract web image features. The Dense SIFT features are extracting the features from visual dictionary.

Bag of Words model is used to create the visual dictionary. In the Bag of visual model, each image is represented as a set of order less visual words. In the visual dictionary label representation, each region of an image becomes a visual “word” of the dictionary. The input image features are extracted by using the Dense SIFT feature extraction algorithm. In the same way the database images features are also extracted.

Finally it compares the input image features with database image feature and it returns the matched features as a result for the query image. The retrieval results are based upon the threshold value.

4.4 Similarity Computation

The fitness function is employed to evaluate the quality of the chromosomes in the population. The use of IGA allows the fusion of human and computer efforts for problem solving. Since the objective of this system is to retrieve the images that are most satisfied to the user needs, the evaluation might simultaneously incorporate user subjective evaluation and intrinsic characteristics of the images. Hence in the Bayesian approach, the quality of a chromosome C with relation to the query q is defined as

$$F(q,c) = \omega_1 . sim(q,c) + \omega_2 . \delta \quad (2)$$

Where sim(q,c) represents the similarity measure between images, δ indicates the impact factor of human's judgment, the coefficients 1 and 2 determine the relative importance of them to calculate the fitness, and Σ i =1. In this paper, they are both set to 0.5. The similarity measure between images is defined as

$$sim(q,c) = \sqrt{\sum_{t\epsilon\{H,S,V\}}(\mu_t{'}^q - \mu_t{'}^c)^2 + \sum_{t\epsilon\{H,S,V\}}(\sigma_t{'}^q - \sigma_t{'}^c)^2} + \frac{H(BM^q, BM^C)}{B\ X\ M} + |E^q - E^C| + \frac{|EHD^q - EHD^C|}{5\ X\ 20} \quad (3)$$

Where $\mu_t{'}$ and $\sigma_t{'}$ represent the normalized meanvalue and standard deviation of the image I in t color space BMI means the image bitmap feature of the image I. EI and EHDI represent the entropy and Edge Histogram Descriptor of the image I. For two images, the hamming distance used to evaluate the image bitmap similarity is defined by

$$H(BM^q, BM^C) = \sum_{j=1}^{m}(IH_j{}^q - IH_j{}^C) + \sum_{j=1}^{m}(IS_j{}^q - IS_j{}^C) + \sum_{j=1}^{m}(IV_j{}^q - IV_j{}^C) \quad (4)$$

A user's preference is included in the fitness evaluated by the user. The impact factor is used to indicate the human's judgment or preferences, and the values of the impact factor are carried out with constant range from 0.0 to 1.0 with an interval of 0.1.

4.5 Retrieval Speed

It is observed that the retrieval speed of the image query based CBIR is faster than the text query based CBIR. To increase the speed of this retrieval system be supposed to separate the image database. The image database can separate the categories like food, flowers, buildings, and beach etc.

5 Results of Retrieval

The Performance of retrieval result is measured by Precision and Recall.

$$\text{Precision} = \frac{\text{No of Relevant Images Retrieved}}{\text{Total No of images Retrieved}}$$

$$\text{Recall} = \frac{\text{No of Relevant Images Retrieved}}{\text{Total No of Relevant images in database}}$$

Table 1. Comparison of Precision Values

Category	Quantization method	Image retrieval system based on Bayesian
Flowers	60	90
Food	59	85
Building	90	93
Beach	80	83
Elephants	52	71
African People	50	77
Horses	68	79
Average	64.85	83.29

Table 2. ComparisonOf Recall Values

Category	Quantization method	Image retrieval system based on Bayesian
Flowers	19	25
Food	11	17
Building	18	18
Beach	16	16
Elephants	10	13
African People	10	12
Horses	13	15
Average	13.88	16.71

6 Conclusion

In CBIR, the low level representation considers the descriptor space and split into multiple regions. It employs unsupervised learning techniques like clustering. The limitation of CBIR is that, it describes using low level features. These features are not able to fill the gap between high-level representation and low-level representation. Hence, the Bag-of-Words model derived from text retrieval system to construct a visual dictionary. In the visual dictionary label representation, each region of an image becomes a visual "word" of the dictionary. The image retrieval is done by considering color, texture, and edge features. The color and bitmap method involves extracting only the local and global features such as mean and standard deviation. But in the proposed technique quantization and DSIFT is applied with Bayesian Algorithm. Thus Bayesian Algorithm is best for Image retrieval.

7 Future Enhancement

The visual features and annotation methods are combined may increase the efficiency of retrieval. The query techniques can be explored in particular domain-specific using

this system. Content based image retrieval can be enhanced by more efficient algorithms than Bayesian Algorithm.

References

1. Antonelli, M., Dellepiane, S.G., Goccia, M.: Design and implementation of Web based systems for image segmentation and CBIR. IEEE Trans. Instrum. Meas. 55(6), 1869–1877 (2006)
2. Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic thumbnail cropping and its effectiveness. In: Proc. ACM Symp. User Interface Software and Technology, pp. 95–104 (2003)
3. Gnaneswara Rao, N., Kumar, V., Venkatesh Krishna, V.: Texture Based Image Indexing and Retrieval. IJCSNS International Journal of Computer Science and Network Security 9(5) (2009)
4. Wang, J.Z., Wiederhold, G., Firschein, O., Wei, S.X.: Content Based image indexing and searching using Daubechie's wavelets- Digital libraries, pp. 311–328 (1998)
5. Lu, T.C., Chang, C.C.: Color image retrieval technique based o ncolor features and image bitmap. Inf. Process. Manage. 43(2), 461–472 (2007)
6. Acharya, M., Kundu, M.K.: An adaptive approach to unsupervised texture segmentation using M-band wavelet transforms. Signal processing 81, 1337–1356 (2001)
7. Apostol, N., Alexander, H., Jelena, T.: Semantic Concept Based Query Expansion and Reranking for Multimedia Retrieval. In: Proc. ACM Int'l Conf. Multimedia, Multimedia 2007 (2007)
8. Silva, S.F.d., Batista, M.A., Barcelos, C.A.Z.: Adaptive image retrieval through the use of a genetic algorithm. In: Proc. 19th IEEE Int.Conf. Tools with Artif. Intell., pp. 557–564 (2007)
9. Delp, E.J., Mitchell, O.R.: Image coding using block truncation coding. IEEE Trans. Commun. COM-27(9), 1335–1342 (1979)
10. Sikora, T.: The MPEG-7 visual standard for content description—An overview. IEEE Trans. Circuits Syst. Video Technol. 11(6), 696–702 (2001)
11. Datta, R., Li, J., Wang, J.Z.: Content-Based Image Retrieval - Approaches and Trends of the New Age. In: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 253–262 (2005)
12. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. Published in Journal of ACM Computing Surveys 40(2) (2007)

Applying Latent Semantic Analysis to Optimize Second-order Co-occurrence Vectors for Semantic Relatedness Measurement

Ahmad Pesaranghader[1], Ali Pesaranghader[2], and Azadeh Rezaei[1]

[1] Multimedia University (MMU), Jalan Multimedia, 63100 Cyberjaya, Malaysia
[2] Universiti Putra Malaysia (UPM), 43400 UPM, Serdang, Selangor, Malaysia
{ahmad.pgh,ali.pgh,azadeh.rezaei}@sfmd.ir

Abstract. Measures of semantic relatedness are largely applicable in intelligent tasks of NLP and Bioinformatics. By taking these automated measures into account, this paper attempts to improve Second-order Co-occurrence Vector semantic relatedness measure for more effective estimation of relatedness between two given concepts. Typically, this measure, after constructing concepts definitions (Glosses) from a thesaurus, considers the cosine of the angle between the concepts' gloss vectors as the degree of relatedness. Nonetheless, these computed gloss vectors of concepts are impure and rather large in size which would hinder the expected performance of the measure. By employing latent semantic analysis (LSA), we try to conduct some level of insignificant feature elimination to generate economic gloss vectors. Applying both approaches to the biomedical domain, using MEDLINE as corpus, UMLS as thesaurus, and reference standard of biomedical concept pairs manually rated for relatedness, we show LSA implementation enforces positive impact in terms of performance and efficiency.

Keywords: Biomedical Text Mining, Bioinformatics, Semantic Relatedness, Latent Semantic Analysis, MEDLINE, UMLS, Natural Language Processing.

1 Introduction

Humans, because of their acquisition of cognitive knowledge, have an ability to effortlessly comprehend contexts (phrases or sentences) which include one or more terms with various meanings. For example, we can semantically distinguish two contexts which both contain the term "*pupil*" from each other by considering this term common appearances with other terms. This type of discernment generally happens through our semantic judgment of relatedness among the involved concepts. An intelligent algorithm that computationally imitates this ability is the goal for many natural language processing (NLP), Geo-informatics, Bioinformatics and Semantic Web studies as this quantification of lexical semantic relatedness is highly critical for variety of their involved applications. Considering these applications, Muthaiyah and Kerschberg [1] used semantic relatedness measures for ontology matching; Pekar et al. [2] employed them to enhance a question answering system in the tourism domain; and Chen et al. [3] applied these measures in machine translation. For the biomedical

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 588–599, 2013.
© Springer International Publishing Switzerland 2013

domain, Bousquet et al. [4] investigated codification of medical diagnoses and adverse drug reactions using semantic relatedness and similarity measures.

The output of a relatedness measure is a value, ideally normalized between 0 and 1, indicating how semantically related two given terms (words) are. Mainly, with respect to the semantic relatedness measurement, there are two models of computational technique available: taxonomy-based model and distributional model.

In taxonomy-based model, proposed measures take advantage of lexical structures such as taxonomies. Measures in this model largely deal with semantic similarity measurement as a specific case of semantic relatedness. For general English text, studies on measuring similarity rely on WordNet, a combination of dictionary and thesaurus, designed for supporting automatic text analysis and artificial intelligence applications. In clinical and biomedical studies, researchers employ the Unified Medical Language System (UMLS), a large lexical and semantic ontology of medical vocabularies maintained by the National Library of Medicine.

The notion behind distributional model comes from Firth idea (1957) [5]: "a word is characterized by the company it keeps". In these measures, words specifications are derived from their co-occurrence distributions in a corpus. These co-occurred features will be represented in vector space for the subsequent computation of relatedness.

Second-order Co-occurrence Vector semantic relatedness measure (also known as Gloss Vector semantic relatedness measure) is a distributional-based approach which has a broad application in text mining and information retrieval. This measure, by constructing definitions (mainly known as Glosses) for the concepts from a predefined thesaurus, estimates semantic relatedness between two input concepts through calculation of the cosine of the angle between those concepts' computed gloss vectors. However, this computed gloss vectors of concepts are impure and rather large in size which often hinders the expected performance of the measure. In this paper, by considering algebraic technique of latent semantic analysis (LSA, also known as LSI), we attempt to enforce some level of elimination on insignificant features, which also results in smaller sizes of the computed gloss vectors. We will show, after a slight change in procedure of Second-order Co-occurrence Vector semantic relatedness measure, by implementing LSA technique as an extra step, we increase the probability of more reliable measurement of semantic relatedness and also improve the final performance of the algorithm in terms of speed for the final relatedness computation.

The remainder of the paper is organized as follows. Section 2 presents semantic similarity and relatedness measures proposed in other studies. In Section 3, we list data and resources employed for our experiments in the biomedical domain for LSA-based Second-order Context Vector measure, and in Section 4, our method for removal of insignificant words using LSA is described. In Section 5, experiments and analysis are given, and in Section 6 the software resources used in the study are presented. Finally, the conclusions and future studies are stated in Section 7.

2 Related Works

The proposed semantic measures in the literature either deal with semantic relatedness or semantic similarity. Similarity between words (to be exact between concept) is

specific case of relatedness between them; as an example "*brooding*" and "*incubation*" are highly similar words but "*brooding*" and "*egg*" are just related. In this study we distinct semantic similarity from semantic relatedness because of their dissimilar descriptions and also their case to case specific applications.

2.1 Semantic Similarity Measures

Measures of semantic similarity are dependent on a hierarchical structure of concepts derived from taxonomies or thesauruses. In this regard, a concept is a specific sense of a word (term), which would denote polysemy. One concept can also have different representative words (terms), which would cause synonymy. The hierarchical structures of the concepts typically get equipped with relations including *is-a*, *has-part*, *is-a-part-of* or any other types of relationship. Semantic similarity measures employ positional information of the concepts in the taxonomy in order to estimate how similar two input concepts are. Figure 1 illustrates part of a Parkinson's Disease (PD) treatment taxonomy.

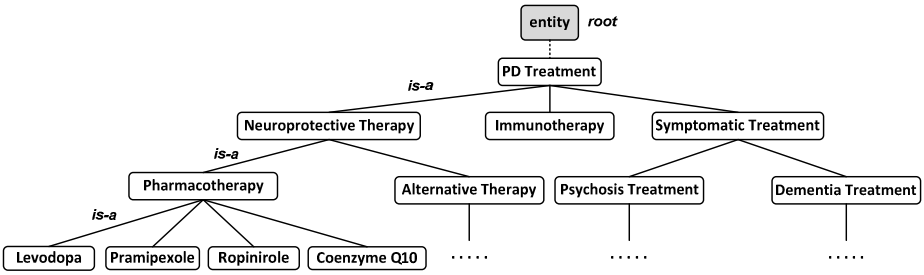


Fig. 1. A portion of the Parkinson's Disease (PD) treatment taxonomy

Path measure [6] is a simple similarity technique based on the calculation of the reciprocal of the shortest path from one input concept to another in the taxonomy. In path measure for finding shortest path between a pair of concepts we count nodes (number of jumps/paths incremented by 1). Path measure does not consider growth in specificity for concepts when we go down from the root towards leaves.

In order to deal with the problem of path measure, Wu and Palmer [7] propose a method which is based on both path and depth of the concepts. By using least common subsumer (LCS) of the two concepts they determined how specific in meaning the two input concepts can be. From information retrieval points of view, when one concept is less frequent than the other one, it is more informative. This informativeness quality denotes abstractness or concreteness in meaning for the concepts, so it is important for semantic similarity measurement. Path and depth based similarity measures lack consideration of this distinctive attribute.

In order to measure how informative one concept is, we need to know frequency of that concept in an external corpus. Resnik introduced this characteristic of a concept as its information content (IC) [8] and quantified it by the negative log of the probability of the concept in a corpus. He also measured two concepts similarity

through calculating IC value of their LCS in a taxonomy. Jiang and Conrath [9] and Lin [10] in their studies improved accuracy of estimated similarity based on concepts' ICs. However, the process of IC calculation for a concept is a challenging task. Absence of the concept's representative word(s) in the corpus and ambiguity of the words in the corpus are just some problematic issues for IC computation.

Pesaranghader and Muthaiyah [11] introduced IC vectors which rely on the definitions of the concepts and their constituent words. They propose that the cosine of the angle between IC vectors of the two input concepts put together and IC vector of their least common subsumer (LCS) is a more reliable and effective way in semantic similarity estimation.

In measures of semantic similarity the dependency on taxonomy relations can be a disadvantage as taxonomies tend to be static and cannot keep up with the rapidly changing structure of knowledge in a given discipline.

2.2 Semantic Relatedness Measures

Measures of semantic relatedness mostly rely on distributional properties of concepts in large text corpora as an easier way to keep track of changes in a given knowledge domain. Nonetheless, Lesk [12] calculate the strength of relatedness between a pair of concepts as a function of the overlap between their definitions by considering their constituent words. Banerjee and Pedersen [13] proposed the Extended Gloss Overlap measure (also known as Adapted Lesk) which augment concepts definitions with the definitions of senses that are directly related to it in WordNet for an improved result. The main drawback in these measures is their strict reliance on concepts definitions and negligence of other knowledge source.

To address forging limitation in the Lesk-type measures, Patwardhan and Pedersen [14] introduced the Gloss Vector measure by joining together both ideas of concepts definitions from a thesaurus and co-occurrence data from a corpus. In their approach, every word in the definition of the concept from WordNet is replaced by its context vector from the co-occurrence data from a corpus and relatedness is calculated as the cosine of the angle between the two input concepts associated vectors (gloss vectors). This Gloss Vector measure is highly valuable as it: 1) employs empirical knowledge implicit in a corpus of data, 2) avoids the direct matching problem, and 3) has no need for an underlying structure. Therefore, in another study, Liu et al. [15] extended the Gloss Vector measure and applied it as Second Order Context Vector measure to the biomedical domain. The UMLS was used for driving concepts definitions and biomedical corpuses were employed for co-occurrence data extraction. In brief, this method gets completed by five steps which sequentially are, 1) constructing co-occurrence matrix from a biomedical corpus, 2) removing insignificant words by low and high frequency cut-off, 3) developing concepts extended definitions using UMLS, 4) constructing definition matrix by results of step 2 and 3, and 5) estimating semantic relatedness for a concept pair. For a list of concept pairs already scored by biomedical experts, they evaluated Second Order Context Vector measure through Spearman's rank correlation coefficient. Figure 2 illustrates the entire procedure.

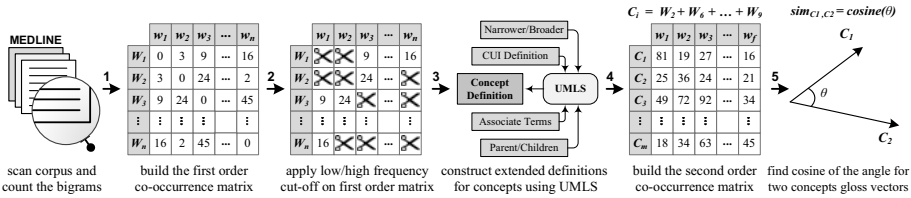


Fig. 2. 5 steps of Gloss Vector relatedness measure

As an example, after stop-words removal and porter stemmer implementation on MEDLINE and then constructing first-order co-occurrence matrix using bigrams frequencies of the rest (changed MEDLINE), low and high frequency cut-off for removing insignificant bigrams from this matrix gets enforced. For the concept *stroke* with the UMLS definition of "*A group of pathological conditions characterized by sudden, non-convulsive loss of neurological function due to brain ischemia or intracranial*", the second order co-occurrence vector will be constructed. For this propose, after removing stop-words and applying porter stemmer on the definition, all available first-order co-occurrence vectors (from cut-off step) of the words appeared in *stroke* definition which are *group*, *pathological, conditions*, *characterized*, *sudden*, *convulsive*, *loss*, *neurological*, *function*, *brain*, *ischemia*, and *intracranial* will be added together. In this sample for the sake of similarity we used only definition of the concept and not its extended definition reachable by appendage of definitions of its directly linked concepts. For other concepts, the procedure of second order co-occurrence vector construction is the same. In order to find the semantic relatedness between two concepts we need to find the cosine of the angle between their computed second order co-occurrence vectors.

Second-order Co-occurrence Vector (or Context Vector) semantic relatedness measure typically tends to generate reliable estimation of semantic relatedness for concept pairs; however, it suffers from two important shortcomings. First, the high and low frequency cut-off phase only considers co-occurrences of the features' (words) frequencies without taking frequencies of individual words into account. This naive cutting approach may cause imperfection of the second order co-occurrence matrix used for final measurement of the relatedness. Second, in many cases, the produced second order co-occurrence matrix is bulky which can hinder usability of this approach in the online and dynamic environments. In this paper, through integrating latent semantic analysis (LSA), we attempt to resolve both of the foregoing drawbacks at the same time.

In the following section we list experimental resources of our experiments in the biomedical domain.

3 Experimental Data

Some of external resources available for the study act as thesauruses; other resources are a corpus to extract required information feeding to the semantic relatedness measures, and a dataset used for testing and evaluation. While MEDLINE Abstract is employed as the corpus, the UMLS gets utilized for construction of concepts extended definitions. The resources worked on in the study are briefly explained in the following subsections.

3.1 Unified Medical Language System (UMLS)

The Unified Medical Language System [1] (UMLS) is a knowledge representation framework designed to support biomedical and clinical research. Its fundamental usage is provision of a database of biomedical terminologies for encoding information contained in electronic medical records and medical decision support. It comprises over 160 terminologies and classification systems, and MedlinePlus Health Topics (MEDLINEPLUS) is one of them. MEDLINEPLUS covers symptoms, causes, treatment and prevention for over 900 diseases, illnesses, health conditions and wellness issues. MedlinePlus health topics are regularly reviewed and get updated daily. The scope of the research is limited to the MEDLINEPLUS vocabulary and all tested concepts belong to this resource. We would employed 2012AB release of the UMLS containing the last version of MEDLINEPLUS with the 954 total concepts (including *root*).

3.2 MEDLINE Abstract

MEDLINE[2] contains over 20 million biomedical articles from 1966 to the present. The database covers journal articles from almost every field of biomedicine including medicine, nursing, pharmacy, dentistry, veterinary medicine, and healthcare. For the current study we used MEDLINE article abstracts as the corpus to build a first order term-term co-occurrence matrix for later computation of second order co-occurrence matrix. We used the 2013 MEDLINE abstract.

3.3 Reference Standard

The reference standard[3] in our experiments is a set of scored medical concept pairs from University of Minnesota Medical School as a result of their experimental study [16]. Eight medical residents were invited for participation in scoring the relatedness of these concept pairs in order to have them voted based on human judgment. Most of the concepts from original reference standard are not included in the MEDLINEPLUS ontology applied in our experiments. Therefore, after removing them from the original dataset, a subset of 39 concept pairs for testing Second-order Co-occurrence Vector semantic relatedness measure was available.

4 Methods

Before explaining our proposed approach for optimizing Second-order Co-occurrence Vector semantic relatedness measure for better performance, we need to know how LSA (also known as LSI) as the heart of our approach works.

4.1 Latent Semantic Analysis (LSA)

LSA was introduced by Landauer and Dumais in 1997 [17]. The method works based on parsing of a collection of different documents. When a set of documents are given

[1] http://www.nlm.nih.gov/research/umls
[2] http://mbr.nlm.nih.gov/index.shtml
[3] http://rxinformatics.umn.edu/data/UMNSRS_relatedness.csv

to the LSA, for a certain term, the algorithm first compute the number of occurrence of that term in each document and then populate the well known term-document matrix. The singular value decomposition (SVD) is then applied to find principal components of this matrix. Considering *A* as our input term-document matrix the result of SVD is as follows:

$$A_{t\times d} = SVD(A) = U_{t\times f} S_{f\times f} V^{t}_{f\times d} \quad (1)$$

The SVD function applied on a matrix, first computes singular values of the matrix (values on *S*'s diagonal) by considering a new latent feature (apart from terms and documents in our case), then produces three matrices based on input matrix and this new feature. The *S* is a rectangular diagonal matrix in which non-negative singular values are sorted downward on the diagonal. By selecting the first *k* indices out of the *f* indices from the produced matrices we will select the *k* principal components. Using *k* for creating new matrices as (2) shows, their multiplication will construct a new matrix that is closely similar to *A*.

$$A_{t\times d} \approx \overline{A}_{t\times d} = U_{t\times k} S_{k\times k} V^{t}_{k\times d} \quad (2)$$

LSA promises by selecting an appropriate *k* the reproduced matrix is highly optimized. Therefore, using cosine approach, this new matrix can be used for any term-term, document-document or term-document relatedness estimation. For the sake of dimensionality reduction, rows of the matrix produced from $U_{txk}S_{kxk}$ can be used for term-term, and columns of the matrix produced from $S_{kxk}V^{t}_{kxd}$ can be used for document-document measurement of relatedness without any need to store the whole new produced matrix.

4.2 Optimized Second-order Co-occurrence Vector Relatedness Measure

In our proposed approach for improving efficiency and performance of the Gloss Vector relatedness measure we have two salient changes in Second-order Co-occurrence Vector relatedness measure procedure represented in Fig. 2: 1) eliminating frequency cut-off phase, and, 2) applying LSA on second order co-occurrence matrix.

In the first move, we ignore low and high frequency cut-offs applied on the first order co-occurrence matrix as the high and low frequency cut-off phase only considers co-occurrences of the features' (words) frequencies without taking frequencies of individual words into account. Therefore, after constructing definitions of the concepts we employ untouched first order co-occurrence matrix (rather than its truncated form after frequency cut-off phase) for the second order co-occurrence matrix construction.

In the second move, after building the second order co-occurrence matrix, by enforcing LSA on this matrix, we will be able to find principal components of this matrix. We supposed that an external data set of concept pairs, already scored in terms of relatedness was available. After examining various temporary *k* principal components from decomposition of the second order co-occurrence matrix for constructing LSA-on-SOC matrices, by computing relatedness for the concept pairs from these matrices (instead of second order co-occurrence matrix in regular form) and comparing the estimated results with human scores of them the best and

permanent *k* principal components for construction of LSA-on-SOC matrix gets selected. It should be mentioned that for constructing LSA-on-SOC matrix, considering (2), we use $U_{mxk}S_{kxk}$ (*m* is the total number of concepts or rows in the second-order co-occurrence matrix) rather than $U_{mxk}S_{kxk}V^{t}_{kxn}$ (*n* is the total number of words or columns in the second-order co-occurrence matrix). This is the main benefit of our approach ending up with valuable amount of reduction of unusable dimensions for the last matrix which gets loaded into memory and then used for relatedness measurement. Furthermore, we believe the produced matrix of LSA-on-SOC is a concept-sense matrix; therefore, each value represents the level of dependency (association) between its corresponding concept (row) and feature/sense (column). As some of these produced values are negative and so carry no logic in terms of dependency (association), we will try to optimize the LSA-on-SOC even more by eliminating these values. We would evaluate the introduced measures and our applied considerations in the next section.

5 Experiments and Discussions

The experiments of the study mainly deal with the relatedness measurement of the considered concept pairs from reference standard using Second-order Co-occurrence Vector measure whether in regular approach or through altered form of this measure with LSA implementation. For the LSA implementation we will represents both results, with and without considering negative values in the LSA-on-SOC matrix.

For the experiment results evaluation, Spearman's rank correlation coefficient for assessing the relationship between the already scored reference standards and the auto-generated semantic relatedness results is employed. Spearman's rank correlation, ρ, is a non-parametric (distribution free) measure of statistical dependence between two variables. Here we assume that there is no relationship between the two sets of data. This algorithm sorts data in both sets from highest to lowest, and then subtracts the two sets of ranks and gets the difference *d*. The Spearman's correlation between the ranks is attainable through formula:

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3}$$

If there are no repeated data values, an exact Spearman correlation (with the value of +1) occurs which means each variable is a monotone function of the other one.

Table 1 represents the result of Spearman correlation for 39 concept pairs in both regular and optimized approach of relatedness measurement and their human scores.

Table 1. Spearman's Rank Correlation of Semantic Relatedness

Second-order Co-occurrence relatedness measure	Spearman's Rank Correlation
Regular approach (high/low frequency cut-off)	0.5145
LSA-based approach (with negative values)	0.5115
LSA-based approach (without negative values)	**0.5279**

In Table 1, the Spearman's rank correlation for the regular approach represents highest result when the finest cut-off points are set. For the LSA-based Second-order Co-occurrence Vector relatedness measure the best generated results are shown when the selected number of principle component is equal to 425 and we have all negative values in the LSA-on-SOC matrix removed or kept. Fig. 3 shows how the computed values of Spearman correlation changes after considering different values of *k* as the selected principle component.

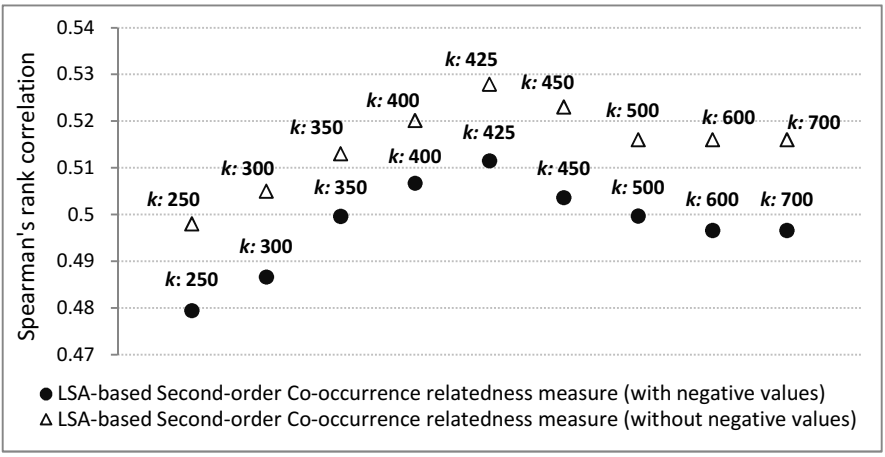


Fig. 3. LSA-based Second-order Co-occurrence relatedness measure Spearman correlation for different values of principle components (k)

In Table 2, we have provided the results of Second-order Co-occurrence relatedness measures on 39 concept pairs considering both frequency cut-off and LSA implementation approaches (in their finest thresholds) along with human judgment of relatedness for the employed concept pairs.

Even though the improved results of Spearman correlation in LSA-based Second-order Co-occurrence Vector relatedness measure seems minuscule, but the main benefit of this approach is the noticeable decrease in the last matrix size. This would increase Second-order Co-occurrence Vector measure's performance, an important issue for online environments. This performance is completely related to the total amount of features (words) in the last matrix. In other words, these differences in the total amount of features in second order co-occurrence matrix in regular form of Second-order Co-occurrence Vector measure and LSA-on-SOC matrix in optimized Second-order Co-occurrence Vector measure is directly proportional to the memory size of these matrices and consequently computation time of relatedness measurement. In the regular Second-order Co-occurrence Vector measure the total amount of features is 31'455 (the second order co-occurrence matrix size is 954×31'455) while in optimized form this amount is 425 (the LSA-on-SOC matrix size is 954×425). It is worth mentioning that these sparse matrices get stored in text files while they just record positive values in the matrix and theirs indices.

Table 2. Different semantic relatedness estimation of 39 concept pairs

Concept 1 (first concept in the concept pair)	Concept 2 (second concept in the concept pair)	Human Scores of Relatedness (range 0-1600)	Gloss Vector (Second-order with frequency cut-off)	Gloss Vector (Second-order with LSA and no negative value)
abortion spontaneous	listeriosis	1005.25	0.6018	0.8187
amyloidosis	arthritis	1127.25	0.7276	0.8824
aneurysm	osteoporosis	275.5	0.2424	0.5601
angina pectoris	atherosclerosis	1357.75	0.6669	0.8427
angina pectoris	diarrhea	553.5	0.483	0.8055
angina pectoris	heartburn	1139	0.3348	0.6363
angina pectoris	cardiomyopathy	1287.25	0.801	0.8951
arthritis	hemochromatosis	857.25	0.7537	0.8761
arthritis	psoriasis	1168.75	0.5971	0.8484
atherosclerosis	heartburn	618.5	0.2613	0.598
atherosclerosis	influenza	416	0.3947	0.7983
atherosclerosis	psoriasis	321.75	0.3624	0.7992
caffeine	plague	637	0.5908	0.8292
coccidioidomycosis	histoplasmosis	1273	0.8882	0.9543
comatose	cataract	330.25	0.7274	0.853
constipation	diarrhea	1023.5	0.69	0.8111
encephalitis	meningitis	1325.75	0.7442	0.8248
epilepsy	cataract	361	0.725	0.8637
flatulence	pancreatitis	333	0.5761	0.724
gastroenteritis	seizures	785	0.6854	0.8279
gonorrhea	syphilis	1432.5	0.8972	0.9438
headache	meningitis	1483.25	0.6801	0.8498
hemorrhoids	infertility	282.25	0.5008	0.7797
hernia	dementia	570.25	0.6395	0.8111
histoplasmosis	mycosis	1272.5	0.8255	0.8908
influenza	pneumonia	1354	0.7319	0.8885
meningitis	mycosis	812.25	0.746	0.875
meningitis	pneumonia	773	0.7526	0.8818
myopathy	myositis	1049.75	0.9311	0.9552
mycosis	pain	958.5	0.6803	0.8537
obesity	snores	1439.75	0.5692	0.8045
osteoporosis	cardiomyopathy	326.25	0.3659	0.6057
pancreatitis	cataract	473.5	0.677	0.8247
pneumonia	septicaemia	1266.25	0.6858	0.8799
psoriasis	dementia	582.25	0.647	0.8883
rabies	acne nos	406.25	0.6845	0.8496
syncope	cardiomyopathy	794.75	0.6229	0.8807
syphilis	dementia	1364	0.809	0.9121
thalassemia	tremor	678.5	0.697	0.8736

6 Software Resources

The software employed for the relatedness measures is part of UMLS::Similarity which is an open source software package which can be downloaded from CPAN[4]. It consists of a suite of Perl modules that can be used in order to calculate the similarity or relatedness between two concepts based on the structure and content of the UMLS. For current study Perl code of UMLS::Similarity and UMLS::Interface (an inter-connector between UMLS database and UMLS::Similarity package) are modified and modules of our methods were included in the package. For LSA computation SVDLIBC[5] is being utilized. SVDLIBC is a C library based on the SVDPACKC library, which is provided by the University of Tennessee.

7 Conclusion

Considering wide applications of Second-order Co-occurrence Vector semantic relatedness measure (also known as Gloss Vector semantic relatedness measure), by applying algebraic technique of latent semantic analysis (LSA, also known as LSI) on the second order co-occurrence matrix as an extra step in this measure, we remove insignificant features of words as a dimensionality reduction phase. More importantly, this slight modification in Second-order Co-occurrence Vector semantic relatedness measure's procedure help to reduce the memory size of the last matrix named LSA-on-SOC matrix used for relatedness measurement, which would lead to higher performance of this measure. In the future works, other mathematical techniques of dimensionality reduction for elimination of invaluable features of words can be examined. Additionally, our proposed approach can be tested on extrinsic tasks such as word sense disambiguation or get evaluated in the other domains of knowledge.

References

1. Muthaiyah, S., Kerschberg, L.: A Hybrid Ontology Mediation Approach for the Semantic Web. International Journal of E-Business Research 4, 79–91 (2008)
2. Pekar, V., Ou, S., Constantin Orasan, C., Spurk, C., Negri, M.: Development and alignment of a domain-specific ontology for question answering. In: Proceedings of the 6th Edition of the Language Resources and Evaluation Conference, LREC-08 (May 2008)
3. Chen, B., Foster, G., Kuhn, R.: Bilingual Sense Similarity for Statistical Machine Translation. In: Proceedings of the ACL, pp. 834–843 (2010)
4. Bousquet, C., Lagier, G., LilloLe, L.A., Le Beller, C., Venot, A., Jaulent, M.C.: Appraisal of the MedDRA Conceptual Structure for describing and grouping adverse drug reactions. Drug Safety 28(1), 19–34 (2005)
5. Firth, J.R.: A Synopsis of Linguistic Theory 1930-1955. In: Studies in Linguistic Analysis, pp. 1–32 (1957)
6. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics 19, 17–30 (1989)

[4] CPAN: `www.cpan.org`
[5] `http://tedlab.mit.edu/~dr/SVDLIBC`

7. Wu, Z., Palmer, M.: Verb Semantics and Lexical Selections. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (1994)
8. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453 (1995)
9. Jiang, J.J., Conrath, D.W.: Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In: International Conference on Research in Computational Linguistics (1997)
10. Lin, D.: An Information-theoretic Definition of Similarity. In: 15th International Conference on Machine Learning, Madison, USA (1998)
11. Pesaranghader, A., Muthaiyah, S.: Definition-based information content vectors for semantic similarity measurement. In: Proceedings of the 2nd International Multi-Conference on Artificial Intelligence Technology (M-CAIT), pp. 268–282 (2013)
12. Lesk, M.: Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice-cream Cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation, New York, USA, pp. 24–26 (1986)
13. Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2002)
14. Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: Proceedings of the EACL 2006 Workshop, Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics together, Trento, Italy, pp. 1–8 (2006)
15. Liu, Y., McInnes, B.T., Pedersen, T., Melton-Meaux, G., Pakhomov, S.: Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In: Proceedings of the 2nd ACM SIGHIT IHI, pp. 363–371
16. Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., Melton, G.: Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. In: Proceedings of AMIA, pp. 572–576 (2010)
17. Landauer, T.K., Dumais, S.T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction and Representation of Knowledge. Psychological Review 104, 211–240 (1997)

A Fuzzy Approach to Multidimensional Context Aware e-Learning Recommender System

Pragya Dwivedi[1,*] and Kamal K. Bharadwaj[2]

[1] Computer Science & Engineering Department
Motilal Nehru National Institute of Technology, Allahabad
pragya.dwijnu@gmail.com
[2] School of Computer and Systems Sciences,
Jawaharlal Nehru University, New Delhi 110067
kbharadwaj@gmail.com

Abstract. Traditional e-Learning recommender systems (EL-RS) based on two dimensions- learners and learning resources, help learners in alleviating information overload by providing suitable learning resources from a potentially overwhelming variety of choices. E-learning recommender systems have received considerable attention in recent years. However, the incorporation of contextual information such as time duration and mood of a learner into the e-Learning recommendation process is still in its infancy. Such contextualization is investigated as an exemplar for EL-RS that can anticipate the learners' requirements. Usually, the representation of learner context is subjective, imprecise and vague. In this paper, we propose a fuzzy approach to multidimensional context aware EL-RS (CA-EL-RS) that includes time duration and mood of a learner as additional dimensions for item based collaborative filtering (IB-CF). The empirical results are presented to demonstrate the effectiveness of the proposed approach in identifying better top N recommendations than traditional IB-CF.

Keywords: e-Learning, Multidimensional Recommender Systems, Context-Aware Recommender Systems, Item-based Collaborative Filtering.

1 Introduction

The unprecedented proliferation of learning resources in e-Learning environment has attracted researchers towards building e-Learning recommender systems (EL-RS) in the recent past for providing appropriate learning resources to learners based on their preferences [1]. In addition to the successful adaptation of recommender systems (RSs) in a wide variety of applications such as movie, travel and news recommendations, the collaborative e-Learning field is also growing popularity for helping learners in alleviating resource overload problem [2]. During the last decade, technology enhanced learning (TEL) community has also extensively employed the recommendation technology for identifying appropriate learning resources from a potentially overwhelming variety of choices [3].

* Corresponding author.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 600–610, 2013.
© Springer International Publishing Switzerland 2013

Today, e-Learning has closely turned into learner-centered for emphasizing pervasive and personalized learning technologies [4]. Learner contextual information consisting of mainly mood, emotion, and time duration for study, place and social interaction can be considered as a valuable factor for the recommendation task. Several attempts have been made in the past for recognizing the importance of contextual information in many disciplines, including e-commerce personalization, information retrieval and data mining. As a result, many researchers have started to turn more on the incorporation of learner contextual information into e-Learning recommender systems (EL-RSs) for improving learning quality.

The incorporation of learner mood play a central role in any learning endeavor and outcomes, especially in online learning. Isen et al.[5] demonstrates that a slight positive mood does not just make you feel a little better but also induces a different kind of thinking in problem solving as well as in decision making. These findings underscore the important effects of emotions on learning. In our work, we considered mood as contextual information of a learner for analyzing its effect on learning activity. Other contextual information, such as time duration and type of content can be considered as additional information for capturing the background information effectively. Time duration can also contribute an important role in deciding the learner mood and interest of a learner on a particular content. For example, in an online learning environment, it is important to determine which learning resources should be recommended to a learner. On weekdays a learner might prefer to read new concepts and prefer to solve reading assignments on weekends. It may be possible that a learner wants to learn a particular topic for a very short period and like to read another topic for a long period based on his mood and interest. In this paper, we present a multidimensional approach that consist of contextual information namely, mood of a learner and time duration for study.

Generally, human decisions employ a wide range of fuzzy terms based on individual perceptions. Fuzziness exists in the expression of mood as well as time duration. In relation to the mood expression, usually people communicate with each other about the mood of a particular person in terms of linguistic expressions such that “Alice is very upset”, “her mood is very bad” and “Bob has jolly mood” etc. whereas time duration can also be expressed in terms of linguistic variables such as “very short”, “short” and “long”. These all contribute to fuzziness that arises from the subjective nature of human. The crisp description of mood and time duration does not reflect the actual case for human decisions and a better approach would be to use linguistic assessments instead of crisp values [6]. Our work in this paper is an attempt towards developing an effective multidimensional context aware e-Learning recommender system (CA-EL-RS) by considering contextual information mainly, mood and time duration as fuzzy concepts. The main contributions of the proposed work are as follows:

- Modelling of learner contextual information using fuzzy sets.
- Development of a multidimensional approach for CA-EL-RS

The rest of the paper is organized as follows: In section 2 we provide background knowledge related to proposed scheme. Section 3 elaborates the proposed framework. In section 4, we describe the experimental setup and criteria to evaluate recommendation quality. Finally, Section 5 provides the concluding remarks and suggests some future research directions.

2 Background

In this section, we firstly discuss about the item based collaborative filtering. Then we explain how context is specified and modeled there.

2.1 Recommendation Techniques

Recommender systems (RSs) have been so popular since the early work in the mid-1990s [6][7]. Collaborative filtering (CF) has been one of the most prevalent techniques in the area of RSs. In CF, the user will be recommended items people with similar tastes and preferences liked in past [8]. Further CF is categorized as user based and item based CF. Our work in this paper is based on item based CF which is described as follows:

Item Based Collaborative Filtering (IB-CF)
The IB-CF explores items in order to identify relations among them [9]. IB-CF first computes the similarity between each pair of items in the system by using various similarity measures like Pearson [10], cosine etc and then selects a set of k most similar items called item neighborhood set for a particular item. Finally, IB-CF predicts an active user preference in a particular item by employing any aggregating function like similarity weighted function. Normally, the rating prediction $Pr^{IB-CF}(u_k, I_j)$ for an item I_j for an active user u_k is computed by the following formula:

$$Pr^{IB-CF}(u_k, I_j) = \frac{\sum_{I \in N_I \cap I_{u_k}} Sim(I, I_j) * r_{u_k, I}}{\sum_{I \in N_I \cap I_{u_k}} |Sim(I, I_j)|} \quad (1)$$

where $Sim(I, I_j)$ represents the similarity between target item I_j and I belonging to a set which has common items between neighborhood set N_I and seen items (I_{u_k}) by the active user and $r_{u_k, I}$ is the rating for the item I_j provided by u_k.

2.2 Description of Context

The notion of context is very broad that has been studied across many research areas, including artificial intelligence, ubiquitous computing, psychology, cognitive science and philosophy. Different disciplines tend to take it according to their own perceptions and what fits their specific goals. One of the most cited definitions of context is described as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" [11]. However, our interest lies in the area of EL-RSs and the context is multifaceted concept, we try to describe the context based on those areas which are related to e-Learning RSs such as information retrieval, e-commerce personalization, data mining and e-Learning. The descriptions of context in these areas are presented in the following Table 1.

Table 1. Description of Context in Different Areas

Area	Description
Information Retrieval	Context is regarded as the set of topics potentially related to the search term [12].
E-commerce personalization	The intent of a purchase made by a customer in an e-commerce application [13].
Data Mining	Those events which characterize the life stages of a customer and that can determine a change in his/her preferences, status, and value for a company [14].
E-Learning	Context is considered as an operational (Network Connectivity, Bandwidth), as a learner preferences (Goal, Knowledge level, Subject domain, Mood), as an environment (Time, Place, Community, Noise levels)

2.3 Contextual Information in Recommender Systems

A novel effort on context-aware recommender systems (CARS) has been made by Adomavicius and Tuzhilin,[15]. The authors presented various approaches where the traditional user/item prototype was extended to support additional dimensions for capturing the context. Further such contextual information can be attained in many ways- explicitly through registration processes, implicitly by automatic detection from environment and through inference mechanism by analyzing of user interactions with tools and resources. Three different approaches are proposed to incorporate contextual information into recommendation process: contextual pre-filtering uses contextual information to filter the dataset before the recommendation process, context post filtering uses context information for adjusting recommendations based on the different contexts and contextual modeling approach employs multidimensional algorithms for generating recommendations using directly contextual information.

In e-Learning, learning activity may also depends on additional contextual information consisting of learning styles, knowledge level, time duration, emotions, mood etc. Therefore, many researchers have developed several context aware ELRSs for enhancing the quality of learning in e-Learning technology. They incorporate different types of context for example, Wan and Okatomato, [1] build a multidimensional e-Learning recommender system considering knowledge level and social interaction as contextual information. Schmidt et al. [16] included a task category while Chen and Kotz ,[17] added time as a context category.

2.4 Multidimensional Model

Initially, the performance of a RS was explicitly dependent on the notion of ratings, the only way of capturing user preferences. For example, in case of e-Learning recommender system, “Alice may assign a rating of 3 (out of 5) for a particular resource (book) entitled *Introduction of Database Management Systems*”. Then, the RS starts

with the initial set (R) of ratings and tries to estimate the rating function (F) for the (learner, resource) pairs that have not been rated yet by the learners

$$F: \text{Learner} \times \text{Resource} \rightarrow R$$

Where R is a totally ordered set of ratings and Learner and Resource are the domains of learners and Resources respectively. Once the function F is estimated for the whole Learner × resource space, a RS can recommend the high-rated resource for learners. On the contrary, we present a multidimensional approach for context-aware e-Learning recommender system, which incorporates learner's contextual information as additional dimensions into the recommendation process. In this regard, ratings are defined with the rating function (F) as:

$$F: \text{Learner} \times \text{Resource} \times \text{Context} \rightarrow R$$

3 Our Proposed Context-Aware e-Learning Recommender System (CA-EL-RS)

In this section, firstly we will describe contextual information in our proposed system and then present modeling of contextual information. Finally, we will discuss the proposed e-Learning recommender system (CA-EL-RS).

3.1 Context Modeling in Our Proposed Model

We consider learner mood and time duration as contextual information for building context aware e-Learning recommender system.

Mood often can be considered as irrational feelings that are beyond our control. However, mood of a learner is complex state of mind and body which is subjective in nature. It consists of physiological, behavioral and cognitive reactions to situations [18]. Consideration of learner mood in learning process plays a vital role because learners' performance during a learning session may be seriously hampered due to mood state. Generally mood is categorized in two classes namely, positive mood and negative mood [19]. But in real life, the intensity of mood also varies. For example, if Bob told to Alice that today I am very happy and Alice said that I am happy. So we can say that Bob and Alice lie in positive dimension of mood and same recommendations will generate based on this dimension. But is this right classification or recommendations (Fig.1). In addition to mood factor, time duration is also important during learning process. Time duration shows how much a learner is interested or not in studying learning resource. It implicitly implies that learning resource is relevant or not for learner. For example if Bob read the learning material for a very short duration it shows different perceptions such "resource is too complex to understand for him/her" or "he/she is in very sad mood so she does not like to learn". Since time duration does not have a sharp boundary i.e uncertainty exists in the representation of time duration. Therefore, we suggest the use of fuzzy sets for the representations of mood intensity and time duration. First of all, mood intensity (Mood) is fuzzified into five fuzzy sets; Extremely Negative (EN), Negative (N), Neutral, Positive (P), Extremely positive (EP), as shown in Fig. 2(a).

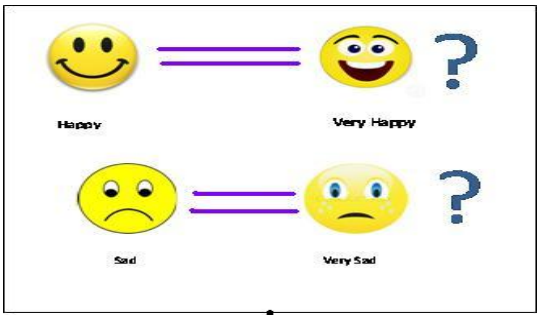


Fig. 1. An example of mood comparisons

Then time duration (TD) is fuzzified into very short duration (VSD), short duration (SD), normal duration (ND), long duration (LD), and very long duration (VLD) as shown in Fig. 2(b).

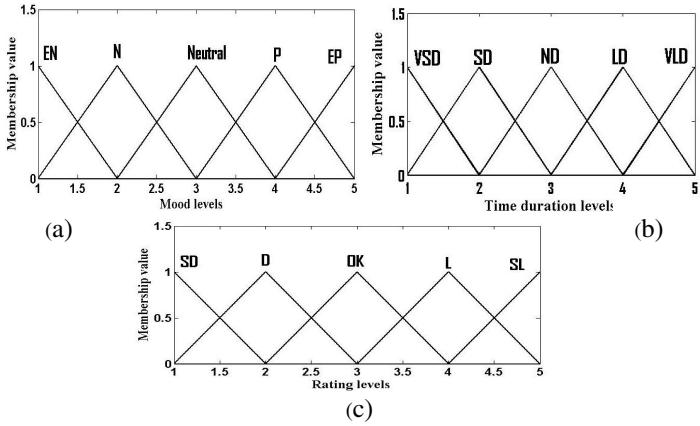


Fig. 2. Fuzzy Levels of (a) Mood (b) Time duration (c) Rating

In this paper, we also considered five fuzzy sets for learner ratings namely strongly disliked (SD), disliked (D), It's OK (OK), liked (L) and strongly liked (SL) as shown in Fig.2(c).

3.2 Proposed Multidimensional Approach

We propose a multidimensional approach for context aware recommender system where an additional dimension is used for generating recommendation model by considering it as virtual resource. Let D be the set of additional dimensions which is defined as $D = \{Mood, Time\ duration\}$. This approach is slightly based on a method proposed by Domingues, et al., [20]. In this approach, we convert multidimensional model (Here 3D) into an extended 2-dimensional model. Main steps involved in this approach are as follows:

- First we generate extended two dimensional models from three dimensional models by considering the third dimension as virtual resource which is illustrated by the following Fig.3 where the values of the additional dimension mood are used as virtual resource.

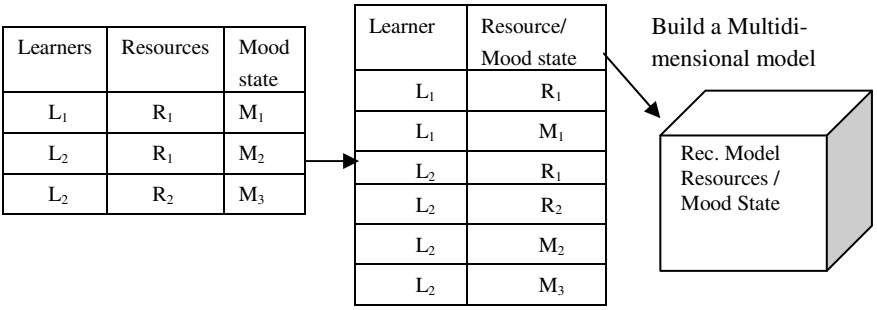


Fig. 3. Multidimensional model transformation

- We compute similarity between all resources as follows: In our approach, it is considered that each resource is represented in terms of learners who have experienced it. Each learner provides his ratings for experienced resources and each rating is classified into five fuzzy sets. In this context, similarity between the resources R_l and R_p, $sim(R_l, R_p)$ is as computed as follows:

$$sim(R_l, R_p) = \left(1 - \frac{\sum_{j \in S} \frac{\sum_{i=1}^{5} \left|\mu_i(r_j^l) - \mu_i(r_j^p)\right|}{5}}{|S|}\right) \tag{2}$$

where S is the set of those learners who have rated these two resources, r_j^l and r_j^p are the ratings provided by j^{th} learner for resources R_l and R_p respectively and $\mu_i(r_j^l)$ and $\mu_i(r_j^p)$ be the membership degrees associated the fuzzy set i for the ratings.

- Finally, we generate the recommendation set as follows: let E_l be the set of experienced resources by an active learner l and O_l is the set of observable resources (Virtual and regular) by a learnerl that is defined as $O_l = E_l \cup d$ where$\forall d \in D$. First we recognize the set C of candidate resources for recommendation such that $c \in C(\text{or IUD} - O_l)$. Then for each c, we compute its recommendation score to the set O_l by using the following formula $R_{c,l}^d = \frac{\sum_{r \in K_c \cap O_l} Sim(c,r)}{\sum_{r \in K_c} Sim(c,r)}$
where K_c is the set of k most similar resources to the candidate resource c.

3.3 Proposed Recommendation Framework

Our proposed context aware e-Learning recommender system (CA-EL-RS) exploits contextual information (mood and time duration) as multidimensional data using

existing traditional recommender systems. We have considered only these pieces of contextual information as virtual resources for building the item based collaborative filtering model (IB-CF). For the sake of simplicity, we have first developed IB-CF framework separately for mood and time duration, then combined them in a weighted manner. Our proposed framework consists of three phases which are defined as:

Phase 1 In this phase, we first consider the mood of learner as a virtual resource for building recommendation model and compute the recommendation score for a candidate resource c using proposed multidimensional approach described in previous subsection which is denoted as $R_{c,l}^{mood}$

Phase 2 Then, we consider the time duration (TD) as a virtual resource for building recommendation model and compute the recommendation score for a candidate resource c using proposed multidimensional approach described in previous subsection which is denoted as $R_{c,l}^{TD}$.

Phase 3 The final recommendation score for a candidate resource can be computed as follows in our approach

$$R_{c,l}^{F} = \delta * R_{c,l}^{mood} + (1 - \delta) * R_{c,l}^{TD} \quad (3)$$

where δ lies between [0, 1].

Finally, the proposed system (CA-EL-RS) selects the top N resources on the basis high recommendation scores and recommends these resources to an active learner l.

4 Experiments and Results

In order to evaluate the performance of proposed scheme, utilizing mood and time contexts as separate dimensions in collaborative filtering framework, we have conducted a set of experiments.

Recently, several researchers raised the issue of dataset in e-Learning recommender system field[2][3]. For the requirement of dataset in e-Learning, RecSys TEL organized workshop in 2010 to discuss and sharing of dataset in TEL community. So due to lack of publically accessible dataset for research [2] in the domain of ELRSs, dataset is generated through the survey in the school of computer and system sciences at Jawaharlal Nehru university, New Delhi. This dataset consist of 50 learners and their ratings provided on 60 resources. For recognition of mood of learners, we generated random numbers for every learner in the range of 1 to 5. Further, in real life, most of classes are scheduled for five or six hours. Therefore, we generated also time duration for each learner on a 5-point rating scale. For our experiments, we choose three subsets from the data, containing 20, 30 and 40 learners called, Sample 1, Sample 2 and Sample 3 respectively. This is to illustrate the effectiveness of the proposed work under varying number of participating learners. Each of the datasets was randomly split into 60% training data and 40% test data. All experiments are run 10 times to eliminate the effect of any bias in the data.

Performance Evaluation

For evaluating the accuracy of recommendations, in this paper, we employ precision, recall, F-measure for evaluating the systems' accuracy of top-N recommendations.

Precision: It is defined as the ratio of the number of selected learning resources to the number of recommended learning resources.

$$\text{Precision} = \frac{\text{Number of relevant learning resources recommended}}{\text{Total number of recommend learning resources}} \quad (4)$$

Recall: Recall is a measure of completeness and can be used as a measure of the ability of our system to all relevant learning resources.

$$\text{Recall} = \frac{\text{Number of relevent learning resources recommended}}{\text{Total number of relevant learning resources}} \quad (5)$$

F- measure: F-measure is the harmonic mean of precision and recall i.e. it means if both values are high then F-measure becomes high.

$$\text{F} - \text{Measure} = \frac{2*\text{precision}*\text{recall}}{(\text{precision}+\text{recall})} \quad (6)$$

4.1 Experiment 1

To demonstrate the feasibility and effectiveness of proposed scheme CA-EL-RS with IB-CF scheme, we compared these schemes via precision, recall and F-measure using equations (4), (5) and (6) in this experiment. The results are presented in Table 2. The precision, recall and F-measure are computed based on the average over 10 runs of the experiments over different datasets. Higher values of precision and recall imply the better performance of the proposed scheme.

Table 2. Precision, Recall and F-measure comparision of CA-EL-RS with IB-CF

Dataset	Scheme	Precision	Recall	F-Measure
Sampe1	IB-CF	0.6283	0.5991	0.6148
	CA-EL-RS	0.6522	0.6446	0.6453
Sample2	IB-CF	0.6311	0.5456	0.5683
	CA-EL-RS	0.5968	0.6494	0.6241
Sample3	IB-CF	0.6213	0.3779	0.4782
	CA-EL-RS	0.6582	0.5376	0.6072

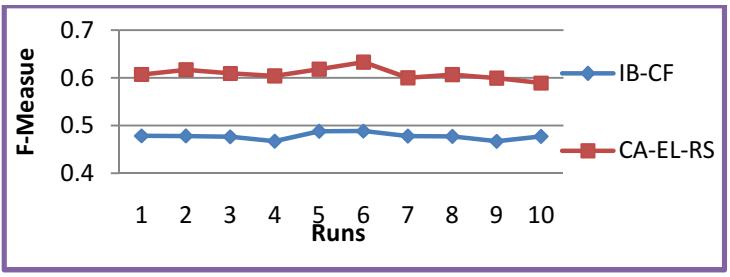


Fig. 4. F-Measure for Sample 3 over 10 runs

Result Analysis

Results presented in Table 2 show the relative performances of IB-CF with our proposed scheme CA-EL-RS. It is clear from Table 2, the proposed scheme CA-EL-RS considerably performed better than IB-CF technique in terms of precision, recall and F-measure. The F-measure for the different runs of the experiment for Sample 3 is depicted in Fig.4. For all the runs, the proposed scheme CA-EL-RS outperforms the IB-CF scheme in terms of F-measure metrics.

5 Conclusion and Future Work

In this paper, we developed an item based collaborative filtering framework utilizing contextual information such as mood and time interval as separate dimensions in addition to learner and learning resource. We proposed the use of fuzzy sets for the representations of mood intensity and time duration for handling vagueness associated with them. To evaluate the effectiveness of our proposed scheme CA-EL-RS, we conducted an experimental study comparing the proposed approach with the traditional IB-CF. Our results indicate that proposed scheme consistently outperforms the IB-CF. Since there is a lack of publically available e-Learning data set for multidimensional recommender system [2]. We have conducted our experiments on synthetic dataset. It would also be important to further challenge the CA-EL-RS approach with new data set when publically available. One of the important further research directions would be to incorporate the idea presented in this work into trust aware e-Learning recommender system [21] [22].

References

1. Wan, X., Okamoto, T.: Utilizing learning process to improve recommender system for group learning support. Neural Comput. & Applic 20, 611–621 (2011)
2. Bobadilla, J., Serradilla, F., Hernando, A.: Collaborative Filtering Adapted to Recommender Systems of E-Learning. Knowledge-Based Systems 22(4), 261–265 (2009)
3. Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., Duval, E.: Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. IEEE Tran on Learning Tech 5, 1–18 (2012)
4. Shen, L., Wang, M., Shen, R.: Affective e-Learning: Using "Emotional" Data to Improve Learning in Pervasive Learning Environment. Educational Technology & Society 12(2), 176–189 (2009)
5. Isen, A.M.: Positive affect and decision making. In: Lewis, M., Haviland, J. (eds.) Handbook of emotions, p. 720. The Guilford Press, Guilford (2000)
6. Kant, V., Bharadwaj, K.K.: Fuzzy computational Models of Trust and Distrust for Enhanced Recommendations. International Journal of Intelligent Systems 28(4), 332–365 (2013a)
7. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual Information in Recommender Systems Using A Multidimensional Approach. ACM Trans. Information Syst. 23, 103–145 (2005)
8. Bharadwaj, K.K., Al-Shamri, M.Y.: Fuzzy Computational Models for Trust and Reputation Systems. Electron Commerce Res. Appl. 8(1), 37–47 (2009)

9. Karypis, G.: Evaluation of Item-Based Top-N Recommendation Algorithms. In: CIKM 2001: Proc. of the Tenth International Conference on Information and Knowledge Management, pp. 247–254, New York, NY, USA (2001)
10. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: An Open Architecture for Collaborative Filtering of Net News. In: Proc. of the ACM Conference on Computer-Supported Cooperative Work (CSCW 1994), pp. 175–186. Chapel Hill (1994)
11. Dey, A., Abowd, G., Salber, D.: A Conceptual Framework and A Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. Hum.-Comput. Interact. 16, 97–166 (2001)
12. Jones, G.J.F., Glasnevin, D., Gareth, I.: Challenges and Opportunities of Context-Aware Information Access. In: International Workshop on Ubiquitous Data Management, pp. 53–62 (2005)
13. Palmisano, C., Tuzhilin, A., Gorgoglione, M.: Using Context to Improve Predictive Modeling of Customers in Personalization Applications. IEEE Trans. Knowl. Data Engineering 20(11), 1535–1549 (2008)
14. Berry, M.J., Linoff, G.: Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Inc., Chichester (1997)
15. Adomavicius, G., Tuzhilin, A.: Context-Aware Recommender Systems. In: Rokach, L., Shapira, B., Kantor, P., Ricci, F. (eds.) Recommender Systems Handbook: A Complete Guide for Research Scientists & Practitioners, pp. 217–250. Springer (2011)
16. Schmidt, A., Beigl, M., Gellersen, H.-W.: There is More to Context Than Location. Computers & Graphics 23(6), 893–901 (1999)
17. Chen, G., Kotz, D.: A Survey of Context-Aware Mobile Computing Research. Technical report, Hanover, NH, USA (2000)
18. Darling-Hammond, L., Orcutt, S., Strobel, K., Kirsch, E., Lit, I., Martin, D.: Emotion and Learning
19. Moridis, C., Economides, A.A.: Mood Recognition During Online Self-Assessment Test. IEEE Transactions on Learning Technologies 2(1), 50–61 (2009)
20. Domingues, A., Jorge, A., Soares, C.: Dimensions as Virtual Items: Improving the predictive ability of top-N recommender systems. Information Processing and Management 49, 698–720 (2013)
21. Dwivedi, P., Bharadwaj, K.K.: Effective Resource Recommendations for E-Learning: A Collaborative Filtering Framework Based on Experience and Trust. In: Proceeding of the International Conference on Computational Intelligence and Information Technology (CIIT 2011), Pune, India, pp. 165–167 (2011)
22. Dwivedi, P., Bharadwaj, K.K.: Effective Trust-Aware E-Learning Recommender System Based on Learning Styles and Knowledge Levels (to appear)

An Analytical Study on Frequent Itemset Mining Algorithms

K. Pazhani Kumar and S. Arumugaperumal

Dept. of Computer Science
S.T. Hindu College, Nagercoil, TamilNadu

Abstract. Data mining is the process of collecting, extracting and analyzing large data set from different perspectives. Fundamental and important task of data mining is the mining of frequent itemsets. Frequent itemsets play an important role in association rule mining. Many researchers invented ideas to generate the frequent itemsets. The execution time required for generating frequent itemsets play an important role. This study yields a detailed analysis of the FP-Growth, Eclat and SaM algorithms to illustrate the performance with standard datasets Hepatitis and Adault. The comparative study of FP-Growth, Eclat and SaM algorithms includes aspects like different support values and different datasets.

Keywords: Frequent Itemset, Mining, Hepatitis, Adult.

1 Introduction

In recent years the size of database has increased rapidly. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data. Data mining refers to discover knowledge in huge amounts of data. It is a scientific discipline that is concerned with analyzing observational data sets with the objective of finding unsuspected relationships and produces a summary of the data in novel ways that the owner can understand and use. Data mining as a field of study involves the merging of ideas from many domains rather than a pure discipline.

The problem of mining frequent itemsets arose first as a sub-problem of mining association rules. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems. The original motivation for searching association rules came from the need to analyze so called supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. Association rules describe how often items are purchased together. For example, an association rules "beer, chips (80%)" states that four out of five customers that bought beer also bought chips. Such rules can be useful for decisions concerning product pricing, promotions, store layout and many others.

2 Problem Study

2.1 Motivation

Studies of Frequent Itemset (or pattern) Mining is acknowledged in the data mining field because of its broad applications in mining association rules, correlations, and

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 611–617, 2013.
© Springer International Publishing Switzerland 2013

graph pattern constraint based on frequent patterns, sequential patterns, and many other data mining tasks. Efficient algorithms for mining frequent itemsets are crucial for mining association rules as well as for many other data mining tasks. The major challenge found in frequent pattern mining is a large number of result patterns. As the minimum threshold becomes lower, an exponentially large number of itemsets are generated. Therefore, pruning unimportant patterns can be done effectively in mining process and that becomes one of the main topics in frequent pattern mining. Consequently, the main aim is to optimize the process of finding patterns which should be efficient, scalable and can detect the important patterns which can be used in various ways [4].

3 Frequent Itemset Mining Algorithms

3.1 FP-Growth Algorithm

One of the currently fastest and most popular algorithms for frequent item set mining is the FP-growth algorithm [3]. It is based on a prefix tree representation of the given database of transactions (called an FP-tree), which can save considerable amounts of memory for storing the transactions. The basic idea of the FP-growth algorithm can be described as a recursive elimination scheme: in a preprocessing step delete all items from the transactions that are not frequent individually, i.e., do not appear in a user-specified minimum number of transactions. Then select all transactions that contain the least frequent item (least frequent among those that are frequent) and delete this item from them. Recourses to process the obtained reduced (also known as projected) database, remembering that the item sets found in the recursion share the deleted item as a prefix. On return, remove the processed item also from the database of all transactions and start over, i.e., process the second frequent item etc. In these processing steps the prefix tree, which is enhanced by links between the branches, is exploited to quickly find the transactions containing a given item and also to remove this item from the transactions after it has been processed.

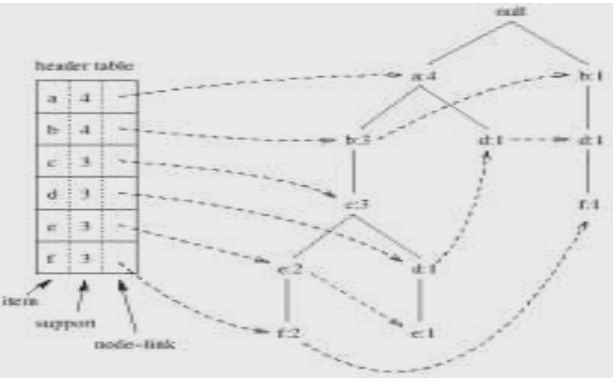


Fig. 1. An Example of FP-Tree

Table 1. An Example of Transaction Database

TID	X
1	{a,b,c,d,e,f}
2	{a,b,c,d,e}
3	{a,d}
4	{b.d.f}
5	{a,b,c,e,f}

The core operation of the FP-growth algorithm is to compute an FP-tree. A frequent pattern tree is a tree structure defined as figure 1.It consists of one root labeled as "root", a set of item prefix sub-trees as the children of the root, and a frequent item header table 1.Each node in the item prefix sub-tree consists of three fields: item-name, count and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none. Each entry in the frequent-item header table consists of two fields, 1. item name and 2. head of node-link, which points to the first node in the FP-tree carrying the item-name.The algorithm FP-tree[4] is as below:

Algorithm 1 (FP-tree construction):

Input: A transactional database *DB* and a minimum support threshold.

Output: Its frequent pattern tree, FP-tree

Method: The FP-tree is constructed in the following steps:

1. Scan the transaction database *DB* once. Collect the set of frequent items *F* and their supports. Sort *F* in support descending order as *L*, the *list* of frequent items.

2. Create the root of an FP-tree, *T*, and label it as "root".

After above process mining of the FP-tree will be done by Creating Conditional (sub) pattern bases:

1 Start from node constructs its conditional pattern base.

2 Then, Construct its conditional FP-tree & perform mining on such a tree.

3 Join the suffix patterns with a frequent pattern generated from a conditional GP-tree for achieving FP-growth.

4 The union of all frequent patterns found by above step gives the required frequent itemset.

In this way frequent patterns are mined from the database using FP-tree.

3.2 Eclat Algorithm

It is a set intersection, depth first search algorithm [5], unlike the Apriori. It uses vertical layout database and each item use intersection based approach for finding the support. In this way, the support of an itemset P can be easily computed by simply intersecting any two subsets Q, R $\subseteq$ P, such that P = Q U R. In this type of algorithm, for each frequent itemset i new database is created Di. This can be done by finding j

which is frequent corresponding to i together as a set then j is also added to the created database i.e. each frequent item is added to the output set. It uses the join step like the Apriori only for generating the candidate sets but as the items are arranged in ascending order of their support thus less amount of intersection is needed between the sets. It generates the larger amount of candidates then Apriori because it uses only two sets at a time for intersection [5]. There is reordering step takes place at each recursion point for reducing the candidate itemsets. In this way by using this algorithm there is no need to find the support of itemsets whose count is greater than 1because Tid-set for each item carry the complete information for the corresponding support. When the database is very large and the itemsets in the database corresponding are also very large then it is feasible to handle the Tid list thus, it produce good results but for small databases its performance is not up to mark.

The Eclat algorithm is as given below [4].

Input: D, σ, i ⊆ I
Output: F [I](D, σ)
1: F [I] :={ }
2: for all I Є I occurring in D do
3: F [I]:= F [I] ∪ {I ∪ {i}}
4: // Create Di
5: Di: = { }
6: for all j Є I occurring in D such that j>I do
7: C: = cover ({i}) ∩ cover ({j})
8: if |C| >= σ then
9: Di: = Di ∪ {(j, C)}
10: end if
11: end for
12: //Depth-first recursion
13: Compute F [I ∪ {i}](Di, σ)
14: F [I]:= F [I] ∪ F [I ∪ {i}]
15: end for

In this algorithm each frequent item is added in the output set. After that, for every such frequent item i, the projected database Di is created. This is done by first finding every item j that frequently occurs together with i. The support of this set {i, j} is computed by intersecting the covers of both items. If {i, j} is frequent, then j is inserted into Di together with its cover. The reordering is performed at every recursion step of the algorithm between line 10 and line 11. Then the algorithm is called recursively to find all frequent itemsets in the new database Di.

3.3 SaM Algorithm

The SaM (Split and Merge) algorithm established by [6] is a simplification of the already fairly simple RElim (Recursive Elimination) algorithm. While RElim represents a (conditional) database by storing one transaction list for each item

(partially vertical representation), the split and merge algorithm employs only a single transaction list (purely horizontal representation), stored as an array. This array is processed with a simple split and merge scheme, which computes a conditional database, processes this conditional database recursively, and finally eliminates the split item from the original (conditional) database.

SaM preprocesses a given transaction following the steps below:

1. The transaction database is taken in its original form. 2. The frequencies of individual items are determined from this input in order to be able to discard infrequent items immediately. 3. The (frequent) items in each transaction are sorted according to their frequency in the transaction database, since it is well known that processing the items in the order of increasing frequency usually leads to the shortest execution times. 4. The transactions are sorted lexicographically into descending order, with item comparisons again being decided by the item frequencies; here the item with the higher frequency precedes the item with the lower frequency. 5. The data structure on which SaM operates is built by combining equal transactions and setting up an array, in which each element consists of two fields:An occurrence counter and a pointer to the sorted transaction (array of contained items). This data structure is then processed recursively to find the frequent item sets. The basic operations of the recursive processing are based on depth-first/divide-and-conquer scheme. In the split step the given array is split with respect to the leading item of the first transaction. All array elements referring to transactions starting with this item are transferred to a new array. The new array created in the split step and the rest of the original arrays are combined with a procedure that is almost identical to one phase of the well-known merge sort algorithm. The main reason for the merge operation in SaM is to keep the list sorted, so that: 1.All transactions with the same leading item are grouped together and 2.Equal transactions (or transaction suffixes) can be combined, thus reducing the number of objects to process.

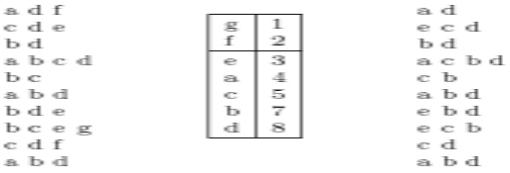


Fig. 2. Transaction database (left), item frequencies (middle), and reduced transaction database with items in transactions sorted accordingly with respect to their frequency (right)

Each transaction is represented as a simple array of item identifiers (which are integer numbers). The transaction list is prepared which are stored in a simple array, each element of which contains a support counter and a pointer to the head of the list. The list elements themselves consist only of a successor pointer and a pointer to the transaction. The transactions are inserted one by one into this structure by simply using their leading item as an index.

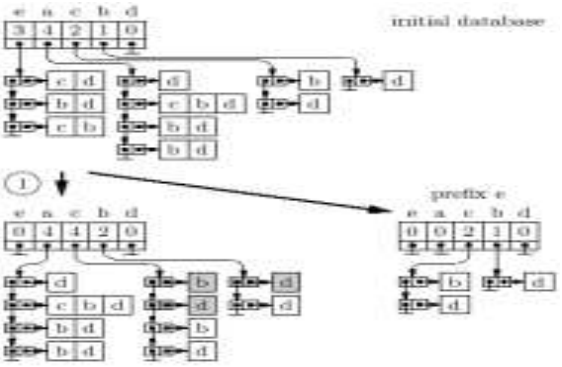


Fig. 3. Procedure of the recursive elimination with the modification of the transaction lists (left) as well as the transaction lists for the recursion (right)

4 Analytical Study

A detailed study has been conducted to assess the performance of the above-said algorithms. The metrics used in the comparison study is the total execution time taken and the number of itemsets generated for different data sets. For this comparison also same data sets were selected as for the above experiment with 30% to 60% of minimum support threshold.

Table 2. Adault data set execution time

SUPPORT	Time in Seconds		
	FPGrowth	Eclat	SaM
30	11.2	10.35	9.25
40	8.30	7.52	6.12
50	5.55	5.20	4.10
60	3.90	3.60	2.80

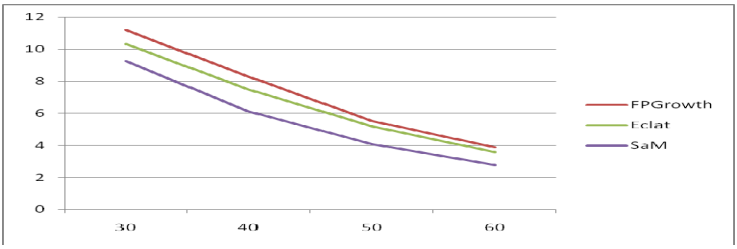


Figure 4 shows that the execution time for the FP-growth,Eclat and SaM algorithms decreases with the increase in support threshold form 30% to 60% for adult dataset FPgrowth takes more time as that compared to Eclat and SaM.

Table 3. Hepatitis data set execution time

SUPPORT	Time in Seconds		
	FPGrowth	Eclat	SaM
30	1.34	1.02	0.9
40	0.71	0.58	0.51
50	0.17	0.12	0.09
60	0.07	0.04	0.03

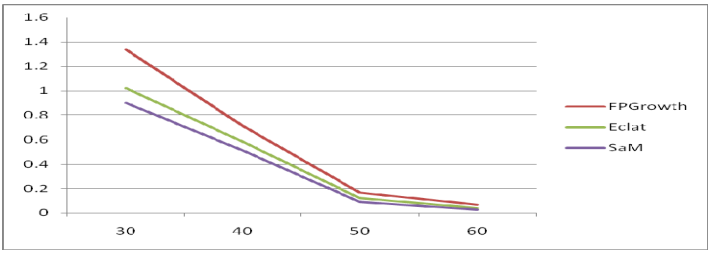


Figure 5 shows that the execution time for the FP-Growth,SaM and Eclat algorithms decreases with the increase in support threshold form 30% to 60% for adult dataset FP-Growth takes more time as that compared to Eclat and SaM.

5 Conclusion

This paper presents the comparative study of three algorithms FP-Growth,Eclat and SaM.This study shows that the SaM algorithm has high performance in various kinds of data, out forms the FP-Growth and Eclat algorithms. The performances of the algorithms strongly depend on the support levels and the feature of the data sets (the nature and the size of the data sets) is observed.

References

[1] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to data mining. Addison Wesley Publishers (2006)
[2] Che, M.S., Han, Yu, P.S.: Data Mining: An Overview from a Database Perspective. Proc. of the IEEE Transactions on Knowledge and Data Engineering 8(6), 866–883 (1996)
[3] Han, J., Pei, H., Yin, Y.: Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD 2000), Dallas, TX, ACM Press, New York (2000)
[4] Pramod, S., Vya, O.P.: Survey on Frequent Itemset Mininbg Algorithms. International Journal of Computer Applications (0975 - 8887)
[5] Borgelt, C.: Efficient Implementations of Apriori and Eclat. In: Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations. CEUR Workshop Proceedings 90, Aachen, Germany (2003)
[6] Borgelt, C.: SaM: Simple Algorithms for Frequent Item Set Mining. IFSA/EUSFLAT 2009 Conference (2009)

Similarity Aggregation a New Version of Rank Aggregation Applied to Credit Scoring Case

Waad Bouaguel[1], Ghazi Bel Mufti[2], and Mohamed Limam[1,3]

[1] LARODEC, ISG, University of Tunis,Tunisia
[2] LARIME, ESSEC, University of Tunis,Tunisia
[3] Dhofar University, Oman
bouaguelwaad@mailpost.tn, belmufti@yahoo.com, mohamed.limam@isg.rnu.tn

Abstract. Credit scoring is one of the most challenging research topics that have been a source of many innovative works in banking field. Choosing the appropriate set of features is one of the most interesting and difficult tasks that have a key effect on the performance of credit scoring models. With the huge amount of feature selection techniques and specially ranking techniques for feature selection, rank aggregation techniques become indispensable tools for fusing individual ranked lists into a single consensus list with better performance. However, in some cases the obtained ranking may be noisy or incomplete witch lead to an unsatisfactory final rank. We investigate on this issue by proposing a similarity based algorithm that extends two standard methods of rank aggregation namely majority vote and mean aggregation based on the similarity between the features in the dataset. Evaluations on four credit datasets show that feature subsets selected by the aggregation based similarity technique give superior results to those selected by individual filters and the standard aggregation techniques.

Keywords: Feature selection, filter, mutual information.

1 Introduction

Financial institutions usually use scoring and consulting system that uses applicants' information to make smart and effective decisions when acquiring new customers. The available information about credit candidate supplies a fundamental element in his credit request acceptation. In one hand, information lack in credit risk valorization is suspected to lead to wrong decision making. In the other hand, tricky and unwanted information may complicate the learning process and also lead to wrong decision. Therefore, choosing the right amount of data in building the scoring model is a fundamental question that should be investigated.

Automating the decision process and manipulate the appropriate set of information, allows credit deciders to make credit decisions and also reduces operational and overhead costs by improving the efficiency and speed of evaluating new credit applicants. In general, the on hand collection of booked loans is used

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 618–628, 2013.
© Springer International Publishing Switzerland 2013

to build a credit scoring model that would be used to identify the associations between the applicant's characteristics and how good or bad is the credit worthiness of the applicant. Generally, portfolios used for scoring task are voluminous and they are in the range of several thousands. These portfolios are characterized by noise, missing values, complexity of distributions and by redundant or irrelevant features [1]. In fact, The more the number of features grows the more computation is required and model accuracy and scoring interpretation are reduced [2,3]. Since the goal in banking instructions is to approximate the underlying function between the input and the target class defining the customer behavior to pay back loan, it is reasonable and important to ignore those input features with little effect on the target class, so as to keep the size of the approximate model small. Thus, removing such features reduces the dimension of the search space and speed up the learning algorithms.

Typically, a feature selection technique looks for a suitable subset of features from the original features set. Feature selection algorithm can be divided into two categories: filter and wrapper. Filter methods use general characteristics of the data independently from the classifier for the evaluation process [4]. The obtained results are generally a ranked list of features, where the top ranked features are the most relevant and features at the bottom are not so relevant or totally unwanted. Wrappers, on the other hand, search for an optimal feature subset by exploiting the resulting classification performance of a specific classifier. Therefore, a wrapper's result is a subset of the most relevant features rather than an ordered list of all the features as given by a filter. Although effective, the exponential number of possible subsets places computational limits on wrapper algorithm which make filter methods more suitable to our study [5].

There are a variety of classical filter methods in previous literature [6,7]. Given the variety of techniques, the question is how to choose the best one for a specific feature selection task?

Since to find the best filtering method is usually intractable in real application, an alternative path is to fuse the results obtained by different filtering methods. In fact, deriving a better rank list from different rank lists, known as rank aggregation, is a hot topic studied in many disciplines. Rank aggregation has gained a huge importance in many data mining applications as credit scoring. However, finding an effective rank aggregation is not always an easy task. In fact standard methods of ranking may not be able to consider similarity between the features in the lists to be aggregated. These methods may give divergent rankings to similar features witch make the final result unacceptable. In this paper, we present a framework to extend previous methods of rank aggregation based on a similarity study applied to credit scoring case. Section 2 presents an overview of rank aggregation and its issues. In Section 3 we propose an extension of rank aggregation using the mutual information between the different features in the lists to be aggregated. Then in Section 4 we present the experimental investigation and finally results and conclusions are given in Section 5 and Section 6.

2 Rank Aggregation

Rank aggregation basically belongs to the ensemble feature selection methods. The concept of ensemble feature selection based feature selectors aggregation was recently introduced in [8]. Ensemble feature selection techniques use an idea similar to ensemble learning for classification [9]. In a first step, a number of different feature selectors are used, and in a final phase the output of these separate selectors is aggregated and returned as the final result a ranked list of features (see Fig 1).

Similar to the case of supervised learning, ensemble techniques might be used to improve the robustness of feature selection techniques. Different feature selection algorithms may yield feature subsets that can be considered local optima in the space of feature subsets, and ensemble feature selection might give a better approximation to the optimal subset or ranking of features. Also, the representational power of a particular feature selector might constrain its search space such that optimal subsets cannot be reached. Ensemble feature selection could help in alleviating this problem by aggregating the outputs of several feature selectors [8]. This concept was especially applied with high dimensional data with few samples [8], [10], but it can be applied to any data dimensionality as it will in the case of credit scoring.

The problem of combining the ranked preferences of many filters is a deep problem. In fact combining a set of lists of ranked features is not always easy as it seems. The rankings provided by the different filters may be in many cases incomplete, or even disjoint. In fact In-complete rankings may come in two forms.

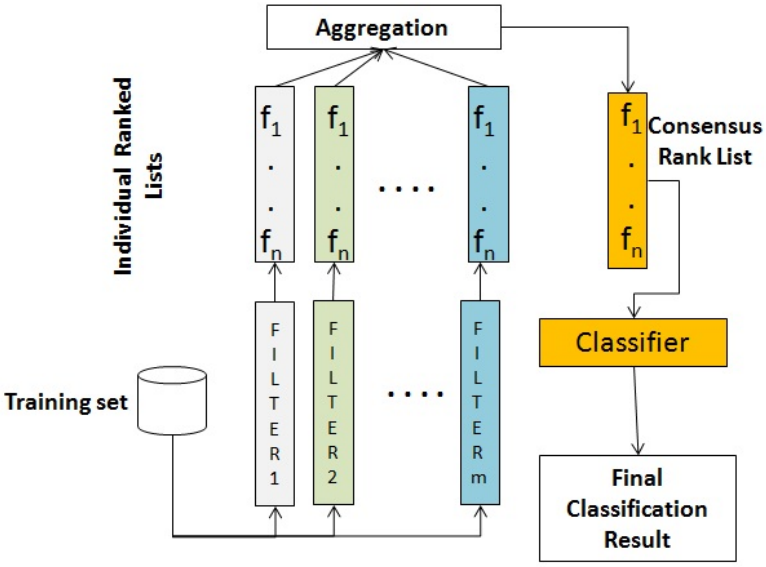


Fig. 1. Rank Aggregation

In the first form, the different filters or some of them may each provide rankings for only the m best features and ignore the remaining features provided in the beginning. In the second form, the used filters may provide complete rankings over a limited subset of available features, due to incomplete knowledge.

Incomplete rankings are common in many financial applications, but still not the only problem with rank aggregation. In fact in the majority of rankings involve a set of similar feature, but despite the similarity between this features they are not ranked similarly which additionally to the problem of incomplete rankings may lead to noisy ranking. Let us give an illustrative example. Assume we have 7 features $\{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$, were f_2 and f_5 are highly similar, but not identical. We consider the two following rank lists from two different filters: list one is given by $\{f_3, f_2, f_7, f_5\}$ and list 2 is given bye $\{f_2, f_7, f_3, f_4, f_1\}$.

If we have no preference of one filter over the other, then standard methods of rank aggregation may interrupt the rankings in the following way: $\{f_2, f_3, f_7, f_5, f_4, f_1\}$. In this standard aggregation the features f_2 and f_5 are given divergent rankings, in spite of their high similarity, which make this kind of aggregation unacceptable. A more acceptable ranking that verify our vision, will be $\{f_3, f_2, f_5, f_7, f_4, f_1\}$. To avoid disjoint ranking for similar features, we present in the next section a simple approach that extend any standard aggregation technique in order to take similarity into account.

3 Proposed Approach: A Rank Aggregation Based on Similarity

We start with a ranked list $Initial_R$ of features $f_1 \dots f_n$, where n is the number of feature in the list $Initial_R$. The handled list can be produced by any standard rank aggregation technique,and where each feature f is included in the original feature set, such that $r(f)$, the rank of a feature f shows the ranking of item f in the ranked list r. Note that the optimal ranking of any item is 1, rankings are always positive, and lower rank shows higher preference in the list.

In each iteration we study the similarity between the first feature and the remaining features in the aggregated list, for that we use the function **SIM**. In this stage the mutual information is chosen as a similarity measure given its efficiency, as opposed to the correlation coefficient measuring only the linear relationship between two random variables. Also mutual information captures nonlinear dependencies. Formally, the mutual information of two continuous random variables X_i and X_j is be defined as:

$$I(x_i, x_j) = \int \int p(x_i, x_j) log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} dx_i dx_j, \tag{1}$$

where $p(x_i, x_j)$ is the joint probability density function, and $p(x_i)$ and $p(x_j)$ are the marginal probability density functions. In the case of discrete random variables, the double integral become a summation, where $p(x_i, x_j)$ is the joint

probability mass function, and $p(x_i)$ and $p(x_j)$ are the marginal probability mass functions.

If the feature in hand have 80% of mutual information with any of the features in the list $Initial_R$, the function **SIM** return 'true' elsewhere it returns 'false'. In case the value 'false' is obtained that means that the feature doesn't have any strong connection with any other features in the list $Initial_R$ and that this feature is in its appropriate place in the aggregated list. In this case this feature is removed from the initial aggregated list and automatically added to the final list. In case the returned value is 'true' we proceed by a set of steps in order to move the similar features closer and resolve the problem of divergent rankings.

In order to make the rank of similar features closer in the aggregated list we take the feature in the top of the list and we study the distance in terms of rank between this feature and the feature with the next rank. We also use the function **PLUS-SIM** to study the distance between the feature with the next rank and the feature with the highest similarity with the first feature. More details are given in Algorithm 1, given below with a detailed description of the different functions used in this approach.

- **SIM(E, L):** return : false, true
  Takes a parameter list L and a feature E and verify if the feature E has a similarity with one of the elements of the list L. If the similarity with one of the elements of the list is superior to 80 %, the function returns true elsewhere false.
- **CONCAT (L, E)**: return : list
  Takes a parameter list L to be concatenated and appends the second argument E into the end of the list L.
- **POS(E,L)**: return: number
  Searches for the feature E in the List L, and returns its position in the list L, or zero if the feature E was not found in L.
- **PLUS-SIM(E, L)**: return : feature
  Searches for a feature in the list L with the biggest similarity to the feature E.
- **SUBLIST(L, P)** : return : list
  Returns a list of the elements in the list L, starting at the specified position P in this list.
- **REMOVE(E,L)**
  Remove the element E given as argument from the list L.
- **DIST-POS(E1,E2,L)** : return : number
  Count the number of position between two given elements E1 and E2 in the list L.
- **PERMUT(E1,E2,L)**
  swap the position of two feature E1 and E2 in the list L.

Algorithm 1.

Require: $Initial_R$: Initial rank aggregation.
Ensure: $Final_R$: Final Rank List.

while $Initial_R = \varnothing$ **do**
  $Var = Initial_R[1]$.
  $Var_{list} = SUBLIST(Initial_R, 2)$.
  if SIM(Var, Var_{list})==FALSE **then**
    $Final_R$= CONCAT ($Final_R$, Var).
    $Initial_R$ =Var_{list}.
  else
    Var_{next}= $Var_{list}[1]$.
    if Var_{next} =PLUS-SIM (Var, Var_{list}) **then**
      $Final_R$= CONCAT ($Final_R$, Var).
      $Final_R$= CONCAT ($Final_R$, Var_{next}).
      REMOVE(Var_{next}, Var_{list}).
      $Initial_R$ =Var_{list}.
    else
      while Var_{next}=PLUS-SIM (Var, Var_{list}) **do**
        if DIST-POS(Var-next ,PLUS-SIM (Var, Var_{list}), Var_{list}) > 1 **then**
          PERMUTE(Var-next, Var,$Initial_R$).
        else
          PERMUTE(PLUS-SIM (Var, $Initial_R$),Var-next,$Initial_R$)).
        end if
      end while
    end if
  end if
end while
Return $Final_R$.

4 Experimental Investigations

4.1 Datasets

The adopted herein datasets are four real-world datasets: two datasets from the UCI repository of machine learning databases (i.e. Australian and German credit datasets), a dataset from a Tunisian bank and the HMEQ dataset.

- Australian: presents an interesting mixture of attributes: 6 continuous, 8 nominal and a target attribute with few missing values. This dataset is composed of 690 instances where 306 are creditworthy and 383 are not. All attribute names and values have been changed to meaningless symbols for confidentiality.
- German: covers a sample of 1000 of credit consumers where 700 instances are creditworthy and 300 are not. For each applicant, 21 numeric input variables are available .i.e. 7 numerical, 13 categorical and a target attribute.
- HMEQ: is composed of 5960 instances describing recent home equity loans where 4771 instances are creditworthy and 1189 are not. The target is a binary variable that indicates if an applicant is eventually defaulted. For

each applicant, 12 input variables were recorded where 10 are continuous features, 1 is binary and 1 is nominal.
- Tunisian: covers a sample of 2970 instances of credit consumers where 2523 instances are creditworthy and 446 are not. Each credit applicant is described by a binary target variable and a set of 22 input variables were 11 features are numerical and 11 are categorical.

4.2 Feature Selection Algorithms

in this study we use aggregation technique on three different filter selection algorithms, Relief algorithm [11], Correlation-based feature selection [12] and Information gain [13]. These algorithms are available in Weka 3.7.0 machine learning package [14].

Relief algorithm evaluates each feature by its ability to distinguish the neighboring instances. It randomly samples the instances and checks the instances of the same and different classes that are near to each other.

Correlation-based feature selection looks for feature subsets based on the degree of redundancy among the features. The objective is to find the feature subsets that are individually highly correlated with the class but have low intercorrelation. The information gain evaluates the worth of an attribute by measuring the number of bits of information obtained for prediction class by knowing the presence or absence of a feature.

4.3 Rank Aggregation Algorithms

Two standard rank aggregation techniques are used in this paper in order to obtain the initial rank list, namely majority vote and mean aggregation.

Majority vote is a common classifier combination method, particularly used in classifier ensemble when the class labels of the classifiers are crisp [15]. In general, majority voting is a simple method that does not require any parameters to be trained or any additional information for the later results [16]. We propose to use majority voting to feature selection in order to fuse an ensemble of filter methods. This method use voting for selecting the features with the major amount of votes. In this case the input is a set of ranking lists generated by several feature selection techniques, and which are sorted in descending order according to their corresponding votes, from the most significant feature to the least one. The output is a single list of features corresponding to the most discriminating features.

Mean Aggregation consists of taking the average rank across all of the ranked feature lists and using that mean value to determine the final rank of the feature. Mean aggregation technique is practical and easy to implement which make it frequently used for ensemble feature selection [17].

4.4 Performance Metrics

In order to evaluate the performance of the proposed algorithm. We use the classification performance of the final chosen features set, obtained respectively

by the majority vote and mean aggregation and their improved version. We used several characteristics of classification performance all derived from the confusion matrix [18]. We define briefly these evaluation metrics.

The precision is the percentage of positive predictions that are correct. The Recall (or sensitivity) is the percentage of positive labeled instances that were predicted as positive. The F-measure can be interpreted as a weighted average of the precision and recall. It reaches its best value at 1 and worst score at 0.

5 Results

For simplicity, each variable is discretized. Then, we split the datasets into a training sample and a test sample, where the first deals with the new feature selection approach and the diverse classification models and the second one checks the reliability of the constructed models in the learning step.Tables 1,2,3 and Table 4 report on the performances achieved by Support Vector Machine (SVM) and Decision Tree (DT) using: the 3 individual filters presented in the previous section, majority vote and mean aggregation, and the improved version of these two aggregation techniques using our approach.

Table 1. Results summary for the Australian dataset

	Precision	Recall	F-Measure
	DT		
Relief	0.923	0.923	0.923
Correlation Coef	0.926	0.924	0.926
Information Gain	0.919	0.944	0.929
Majority Vote	0.926	0.946	0.934
Similarity aggregation	0.934	0.950	0.942
Mean aggregation	0.927	0.934	0.931
Similarity aggregation	0.930	0.942	0.936
	SVM		
Relief	0.941	0.880	0.909
Correlation Coef	0.931	0.860	0.905
Information Gain	0.910	0.908	0.890
Majority Vote	0.931	0.908	0.910
Similarity aggregation	0.941	0.911	0.926
Mean aggregation	0.931	0.890	0.910
Similarity aggregation	0.943	0.908	0.925

From Table Tables 1,2,3 and Table 4 we notice that the individual filters do not give the best classification performance because they may not always give the best set of features. There is obviously a strong similarity in the feature sets selected by different approaches. A more detailed picture of the achieved results shows that the precision of the two aggregation techniques is better than the individual filters. Consistent with the theoretical analysis for feature selection, the aggregation approach usually out-performs single filters.

Table 2. Results summary for the German dataset

	Precision	Recall	F-Measure
	DT		
Relief	0.692	0.511	0.588
Correlation coef	0.721	0.500	0.591
Information Gain	0.750	0.580	0.654
Majority Vote	0.781	0.586	0.658
Similarity aggregation	0.790	0.595	0.679
Mean	0.781	0.586	0.656
Similarity aggregation	0.788	0.596	0.678
	DT		
Relief	0.694	0.489	0.573
Correlation coef	0.705	0.489	0.577
Information Gain	0.738	0.545	0.627
Majority Vote	0.766	0.557	0.645
Similarity aggregation	0.781	0.563	0.654
Mean aggregation	0.766	0.552	0.627

Table 3. Results summary for the HMEQ dataset

	Precision	Recall	F-Measure
	DT		
Relief	0.819	0.836	0.81
Correlation coef	0.838	0.974	0.901
Information Gain	0.819	0.836	0.81
Majority Vote	0.853	0.976	0.912
Similarity aggregation	0.859	0.981	0.916
Mean	0.850	0.966	0.904
Similarity aggregation	0.857	0.982	0.915
	SVM		
Relief	0.845	0.807	0.728
Correlation Coef	0.822	0.828	0.784
Information Gain	0.822	0.828	0.784
Majority Vote	0.835	0.989	0.905
Similarity aggregation	0.840	0.990	0.909
Mean aggregation	0.830	0.987	0.902
Similarity aggregation	0.836	0.991	0.907

The results in Tables 1,2,3 and Table 4 show that in this experiment, the obtained feature set by the rank aggregation with similarity, without fail outperformed the results from the features by the standard aggregation methods namely majority vote and mean aggregation, as measured by both selected precision, recall and F-measure. This experiment serves as a real world example of noisy rankings, with a mixture of characteristics of partial and top-k lists, which benefits from the addition of similarity information to rank aggregation.

Table 4. Results summary for the Tunisian dataset

	Precision	Recall	F-Measure
	DT		
Relief	0.827	0.847	0.830
Correlation coef	0.833	0.850	0.832
Information Gain	0.822	0.852	0.826
Majority Vote	0.866	0.985	0.921
Similarity aggregation	0.83	0.992	0.929
Mean aggregation	0.875	0.964	0.917
Similarity aggregation	0.878	.0972	0.923
	SVM		
Relief	0.722	0.85	0.781
Correlation coef	0.769	0.847	0.784
Information Gain	0.851	0.994	0.917
Majority Vote	0.849	0.999	0.930
Similarity aggregation	0.860	0.997	0.923
Mean aggregation	0.840	0.994	0.927
Similarity aggregation	0.857	0.999	0.922

6 Conclusion

Feature selection has always been an important step in credit scoring process. With the important number of feature selection techniques, rank aggregation was proposed as a smart solution combining the results of many filtering methods. Despite their efficacy, aggregation techniques may in some cases suffer from noise. In this paper we present an algorithm for improving feature rankings. The method relates the similarity of the feature to their ranking. Here we offer a promising aggregation approach for feature selection, there are still aggregation possibilities that need to be discussed. Further work is needed to expand our algorithm to other similarity measures.

References

1. Piramuthu, S.: On preprocessing data for financial credit risk evaluation. Expert Syst. Appl. 30, 489–497 (2006)
2. Liu, Y., Schumann, M.: Data mining feature selection for credit scoring models. Journal of the Operational Research Society 56, 1099–1108 (2005)
3. Howley, T., Madden, M.G., O'Connell, M.L., Ryder, A.G.: The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowl.-Based Syst. 19, 363–370 (2006)
4. Forman, G.: BNS feature scaling: an improved representation over tf-idf for svm text classification. In: CIKM 2008: Proceedings of the 17th ACM Conference on Information and Knowledge Mining, pp. 263–270. ACM, New York (2008)
5. Wu, O., Zuo, H., Zhu, M., Hu, W., Gao, J., Wang, H.: Rank aggregation based text feature selection. In: Web Intelligence, pp. 165–172 (2009)

6. Wang, C.M., Huang, W.F.: Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. Expert Syst. Appl. 36, 5900–5908 (2009)
7. Bouaguel, W., Bel Mufti, G.: An improvement direction for filter selection techniques using information theory measures and quadratic optimization. International Journal of Advanced Research in Artificial Intelligence 1, 7–11 (2012)
8. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
9. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
10. Schowe, B., Morik, K.: Fast-ensembles of minimum redundancy feature selection. In: Okun, O., Valentini, G., Re, M. (eds.) Ensembles in Machine Learning Applications. SCI, vol. 373, pp. 75–95. Springer, Heidelberg (2011)
11. Kira, K., Rendell, L.: A practical approach to feature selection. In: Sleeman, D., Edwards, P. (eds.) International Conference on Machine Learning, pp. 368–377 (1992)
12. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the Seventeenth International Conference on Machine Learning, pp. 359–366. Morgan Kaufmann (2000)
13. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. (1993)
14. Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D.: Weka manual (3.7.1) (June 2009)
15. Kuncheva, L.I., Bezdek, J.C., Duin, P.W.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 34, 299–314 (2001)
16. Guldogan, E., Gabbouj, M.: Feature selection for content-based image retrieval. In: Signal, Image and Video Processing, pp. 241–250 (2008)
17. Wald, R., Khoshgoftaar, T.M., Dittman, D.J.: Mean aggregation versus robust rank aggregation for ensemble gene selection. ICMLA (1), 63–69 (2012)
18. Okun, O.: Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations. In (2011)

Learning a Concept Based Ranking Model with User Feedback

E. Umamaheswari and T.V. Geetha

Department of Computer Science and Engineering
Anna University, Chennai - 600016, Tamilnadu, India
vasanthuma28@gmail.com, tv_g@hotmail.com

Abstract. Automatically learning a ranking model is becoming an essential task for effective information retrieval. Its advantage lies in the ability to combine the user feedback with the conceptual features. However, learning to rank methods require large, training and test data sets, to gather feedback from the user. Existing learning-to-rank methods focus either on user feedback or at the document level or on query-dependent feature scores. In this paper, we consider the implicit and explicit feedback from the user using a set of test data. This test data is given as input to the training phase, in which we find the document conceptual and query-dependent document features. Hence, the impact of the user feedback on the search query is identified by learning the document, and query level conceptual features, using statistical methods. A DCG (Discounted Cumulative gain) score is calculated and used in comparing the ranking in subsequent iterations. The learning process continues until there is no change in the DCG score. The learned ranking model is compared with our base ranking method and achieved 20% improvements in nDCG score.

1 Introduction

A good search engine intuitively requires high relevant results at the top 10 retrieved documents. Fine tuning the search results by adjusting some ranking parameters is important in any Information Retrieval system. Any ranking model learned by a machine learning algorithm, is termed as learning to rank. The aim of this work automatically learns a ranking model from the training corpus. This training corpus consists of a list of ranked objects, and associated query set from the base search system. Learning to rank [1] can be classified into point−wise, pair−wise and list wise approaches based on their input and loss function. The Point wise approach[2] takes only a single query and document pair to predict the best score.

All supervised machine learning algorithms use the point wise approach. The Pair wise approach[3] handles the ranking problem as a binary classification, by comparing a pair of documents to the given query. However, in the case of the List wise approach[4] the input space contains the list of objects to be ranked, and its associated permutations. It finds the loss function for different permutations of a ranking model. User feedback is important for good learning of a

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 629–639, 2013.
© Springer International Publishing Switzerland 2013

ranking model. Radlinski (2007) [5] elaborates his work on click through data, and proves the impact of the user feedback in the quality of the information retrieval. User feedback can be retrieved implicitly and explicitly. Explicit user feedback is retrieved by manual user rating of the document. Implicit user feedback can be retrieved automatically, by using user search logs, such as click through, time spent on the page, Number of users for a page, etc. All the previous approaches[5] [6] consider only the user click-through rate as back. In our method we additionally consider time spent on the page and the Number of users per page as implicit feedback.In the context of considering document features in the learning to rank model, previous approaches[7] focus more on the relevance of the document with respect to frequency of occurrence of the term. It ignores the conceptual features with respect to the user perspective.

Instead, in our approach we consider the conceptual features of the query and document, using the Universal Networking language (UNL) [8]. In this approach, the user implicit and explicit feedback and its conceptual feature are learned iteratively. At each iteration, the base ranking function is modified with the help of user feedback, and its conceptual features. A Discounted Cumulative Gain (DCG) score is computed at each iteration, and compared with those of subsequent iterations. The learning stops when there is no change in DCG score of the ranking model. Here, the UNL acts as an intermediate representation [8] which is designed to extract semantic data from the natural language text. It acts as an Interlingual framework, that converts the source language text into UNL representation. It helps to express the semantics at the word, sentence, and document levels. The natural language text from the document is processed sentence by sentence, and represented as a set of the Universal words (UWs) and relations (link between UWs). Word level semantics is expressed [9] by UNL constraints, and context level information is conveyed by using UNL attributes.

The representation of UNL is given in Fig. 1; and the extracted indices from the UNL graph is given in Table 1.The rest of the paper is organized as follows. Section 2 elaborates the related work. Section 3 describes the methodology. Section 4 discusses the critical evaluation of the results. In section 5 we discuss the conclusions and future work.

2 Related Work

Many research works have been proposed recently to learn the ranking function. In this section, we focus on learning the ranking function based on semantic features and user feedback. Learning to rank using semantic features [10] is proposed with the SVM classifier, using pair wise training data. In order to extract the semantic features from a document, it uses modifying relations to extract the semantics of the word in the document. Moreover, defining the modifying relations itself requires a huge knowledge base, and it depends on the predicate of the subject word in the sentence. Finding the semantic relation between terms may lead to wrong interpretation of the terms, which are not available in the knowledge base.

Learning-to-rank with a lot of word features, proposed by Bing Bai (2010)[7], presents supervised semantic indexing, which considers only synonymy and polysemy. Though this approach considers semantic relationship and can be useful across different languages it does not consider the conceptual relationship between words in the document. Moreover, it does not consider user feedback. Other statistical approaches for learning to rank[11] have computational complexity when dealing with huge data, and require more permutations and combinations. This approach considers only the word level distribution of terms and their semantics across documents.

Learning to rank requires a sufficient amount of training data. Many methods are proposed to improve the quality of the training data. One such method proposed by Jingfang Xu (2010)[12], automatically detects the judgment errors, using click-through data. The Optimization of the web search engine using web click-through data [13], proves that the document which is visited by more users, is more important than the others. It fails to give importance to the document which has a low click. In our approach, though the document has a low click we consider additional features, such as the userś explicit rating and its conceptual features.Na Chen and Viktor K. Prasanna (2011) [14] propose a learning-to-rank method, which takes complex semantic relationships, and simultaneously considers user preferences. However, this approach improves the ranking quality, by capturing user preferences and semantic relationships. It considers only ontology like semantic relationship and the explicit ranking given by the user. Moreover, it is personalized, and does not address the interactive user feedback. However, our approach for learning-to-rank extracts both implicit and explicit feedback from the user. Explicit user feedback from different users may vary, and this is not sufficient to find the importance of the document to the query. Hence, in order to predict the user feedback errors, we additionally consider user click-through information. Since many learning to rank algorithms require huge training sets, performance is the major bottle neck when we rank the ranking model online. We have also addressed the performance-related issue, by adopting learning to rank algorithm offline. There are only limited methods, which consider both semantic relationships, and user feedback. One such method proposed by Nachan (2011) which learns semantic features using a statistical model, simultaneously learns the user preferences which are personalized. Our work differs in taking conceptual features by using UNL semantics, and we consider both implicit and explicit feedback, and our ranking model is not personalized. Similarly Yajuan Duan et al, (2010) adopt learning to rank for tweets, by using the popularity of tweets, content relevance and account authority. Since tweets have limited content, the use of limited features for learning a ranking function for tweets is sufficient, but in the case of tourism based search engine, it requires additional parameters to learn the ranking model. Hence, in our approach, the ranking model is learned iteratively, by comparing the DCG score at each iteration. It takes explicit and implicit user feedback, along with the conceptual features, with respect to document-document and document-query. Simultaneously considering both user feedback and conceptual features can improve the quality of the training data. Moreover, the training data is taken by

analysing 10 different query categories of the query set from the tourism domain, with the help of linguistic experts. The proposed approach also incorporates different user feedback, which helps to find relevance of the document which has a low click. Hence, we concentrate both on improving the quality of the training data and the ranking model.

3 Methodology

The Learning to rank method described in this paper modifies the ranking parameters of the previously developed multilevel UNL based ranking model[15] and it is explained below.

3.1 Base Ranking Model

The Universal Networking Language (UNL) based semantic search sub-system in CLIA[15] aims at retrieving documents containing not just the key-words in the query, but documents that contain conceptually and semantically related information. The system describes an UNL based Conceptual approach to search the Tamil documents. The documents are represented using UNL graphs and index structures[16] are built separately for Concept Relation Concept (CRC), Concept Relation (CR) and Concept (C) components of the UNL graph. The user query is converted into an UNL graph after query expansion. The query expansion is based on the context of the query determined by analyzing the conceptual indices. Searching and Ranking are primarily based on CRC, CR and C matches and additional linguistic, statistical, sentence and expanded query term information. The UNL representation of a Tamil sentence and its corresponding CRC, CR, C indices is shown in Figure 1. The UNL based search system is the base system which is intended to provide advanced level search based on semantics. It consists of UNL based input (query) processing[17], document processing[9], indexing[16], searching and ranking[15]. Specifically, the input query is expanded, and enconverted to a UNL sub graph; important components of documents are enconverted into UNL graphs; indexing is based on UNL concepts and relations; searching and ranking is based on the match between the query UNL sub graphs and the document UNL graphs. List of indices extracted from UNL graph is given below.

1. CRC (Concept−relation−concept) Indices
   (a) Mahabalipuram (icl>place) − plc − is(aoj>thing)
   (b) Is (aoj>thing) − plc − Famous (aoj>thing)
   (c) Famous (aoj>thing) − mod − Sculpture (icl> fine arts)
   (d) Historical (aoj>thing) − mod − Sculpture (icl> fine arts)
2. CR(Concept−relation) Indices
   (a) Mahabalipuram (icl>place) − plc
   (b) is (aoj>thing)−plc
   (c) Famous (aoj>thing) − plc

மாமல்லபுரம் வரலாற்றுச் சிறப்புள்ள சிற்பங்களுக்கு பெயர்பெற்றது.

mamallapuram varalaatruch chirpangalukku peryarpetrathu

Mamalapuram is famous for historical sculptures.

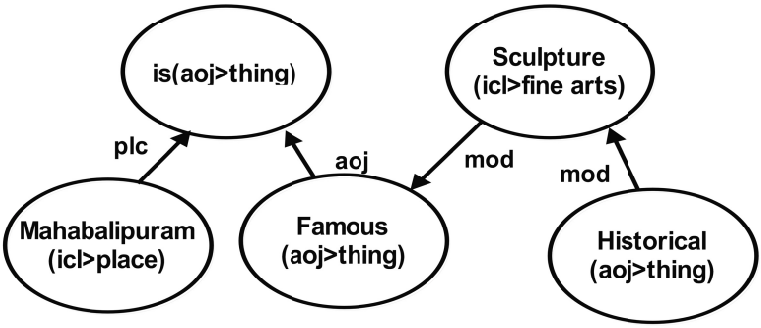


Fig. 1. Representation of a Tamil Sentence in UNL

(d) Famous (aoj>thing) − mod
(e) Sculpture (icl> fine arts) − mod
(f) Historical (aoj>thing) − mod

3. C(Concept) Indices ...
   (a) Mahabalipuram (icl>place)
   (b) is (aoj>thing)
   (c) Famous (aoj>thing)
   (d) Sculpture (icl> fine arts)
   (e) Historical (aoj>thing)

The basic searching procedure is based on the complete CRC Match, or partial CR or C matches between the query sub graphs and the corresponding UNL index sub−graph. The ranking is based on the Degree of match (T_a), Concept association (T_b) and Index based feature weight (W_I).

Three Level Ranking Level 1: Degree of match categorization, prioritizes documents based on complete match ($CRCmatch$), partial $CRmatch$ or $Conceptonly(C)$ matches. Level 2: Concept association categorization is based on whether the match is a term match, concept match or expanded concept match. Level 3: Ranking is based on index based features, such as frequency of occurrence of the term and concept in the document, position weight, Named Entity (NE) weight and Multiword (MW) weight. Hence, the base ranking model is defined as follows.

$$W_{baseRankModel} = T_a.T_b * W_I \tag{1}$$

The UNL based ranking model first ranks the document based on the Degree of match (Ta) between the query UNL subgraphs and the document UNL subgraphs, which results in a different set of documents based on CRC match, CR

match and C Match. These documents are then ranked based on the concept association features, such as term match or concept match. These results at the third level, are ranked based on the index based feature weight (WI). The base multilevel UNL based search and rank[15] function is modified by additionally including user feedback and their conceptual feature score, with the Index based feature weight (W_I). The steps for learning the ranking model are given below.

4 Learn-to-Rank

The training data for learning to rank can be found in two ways; either by using human judgement or by the search log of the user. In this approach, a set of queries is randomly selected from the query log into a search system. The query log is chosen from 10 different categories related to tourism domain such as <Tourist Place>, <Festival>, <Special about a place>, <Food available>, <Temple>, <Habitual>, <Entertainment>, <Cities>, <Facilities> and <Naturals> so that the learning will be diverse and the quality of learning data will be high. In order to retrieve the training data a set of queries with different category are given as input to the UNL based search sub system [15], and as a result of this each query is associated with the documents. Human a judgement is then assigned to make relevance judgements on all the query document pairs. Relevance Judgements are usually rated at five levels; for example, perfect, excellent, good, fair, and bad. In the case of multiuser feedback, the previous approaches assign a majority voting for a given query. But the user perspective is not consistent for different communities of users, and requires domain experts to judge the correct ranking. The implicit feedback is calculated by storing the user click-through information and time spend on the page. As mentioned earlier, the relevance judgements are collected manually from different users, and are in the form of a 5 scale rating, from the most relevant to the least relevant, with respect to a given query. We denote this information as follows: For a given set of queries and its top relevant documents, the user feedback is calculated manually. Therefore, the test data contains the query set $Qs = q_1, q_2 \cdots q_n$ and its associated document collection D_N, where N varies from 1 to N. The base ranking function[15] is modified by incorporating the user implicit and explicit feedback, along with its conceptual features with respect to the query and document. Hence, the ranking function is defined as follows.

$$f(q_i, D_{ij}) = W_{baseRankModel} * Learned - weight \tag{2}$$

In the above ranking function $W_{baseRankModel}$ is modified by adding $learned_{weight}$ $((q_i, D_{ij}))$.(q_i, D_{ij}) defines the weight associated with the conceptual features and their user feedback.

Steps:

– For a given set of queries q_i, collect the documents with the user relevant feedback (both explicit and implicit)

– Read the document UNL graph to learn the conceptual feature weight.
– Modify the rank
– Compute the DCG score
– Compare the DCG with the previous iteration.
– If changes are found, repeat from step1.
– If no changes are found, stop the process.

The process flow is given in the Fig.2.

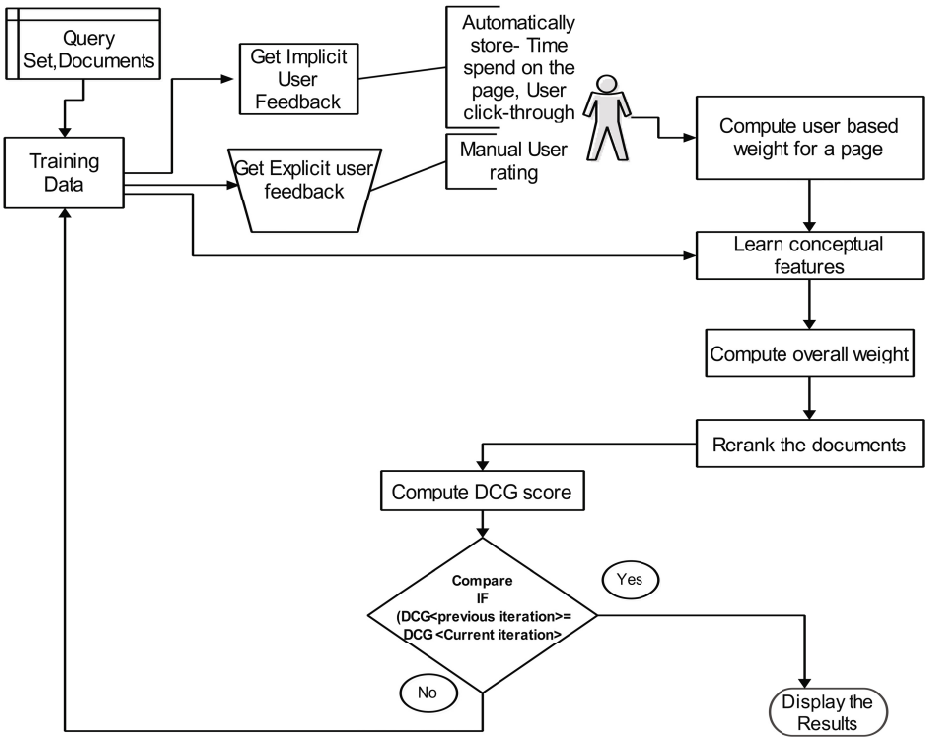


Fig. 2. Process flow of concept based learn to rank

4.1 Retrieval of Implicit and Explicit Feedback from the User

User behaviour can be analysed by the implicit and explicit user feedback, given by the user. Each user assigns a score manually for a set of queries and its top 20 documents. Implicit user feedback is measured by using a five scale rating scheme. In this rating system, users rate the URL for a given query as 5, if it matches exactly with the search query, and the whole document itself talks about the query. They will rate it only as 4 if a few lines or a small portion of the document talks about the query. It is assigned 3 for moderately relevant

results, and 2 or 1 if only a link is relevant and a single line talks about the query. Similarly explicit feedback such as Time spends on the page, Number of click-through; Total number of users for that page are assigned, and a user's feedback score is assigned on the first level. Therefore, user feedback log contains query, document collection with user rating. This log is updated periodically and stored in the Lucene index for future reference. Hence, the storage structure User feedback log is given in Table 2. The user feedback file contains the following parameters to learn the implicit and explicit feedback of the user.

- User ID
- Query
- Implicit and Explicit User Feedback

The user feedback score is computed as follows.

$$\lambda(d,q) = \frac{C^T(d,q,click)}{T^C(d)} * \frac{U^R(d,q)}{MaximumRate} * \frac{T^S(d,q)}{60} * \frac{N^U(d,q)}{T^U(d)} \quad (3)$$

$\lambda(d,q)$ represents the implicit and explicit feedback to the user.
$\alpha(d,q)$ represents the document specific feature score assigned by the base ranking model.
$C^T(d,q,click)$ represents the number of click-through for a given query and its document.
$U^R(d,q)$ represents the five scale rating assigned by each user.
$T^S(d,q)$ represents the time spent on the page for a given query and its document
$N^U(d,q)$ represents the number of users for a given query and its document.
Hence, $\lambda(d,q)$ is computed for top 20 urls and stored in the index to extract conceptual features.

4.2 Computation of Conceptual Feature Score

During this stage, the conceptual features are extracted by examining the documents UNL sub graphs. Usually, the document which has a high user feedback score is more important than the document with low user feedback. We shall first learn the conceptual features of the high user feedback.Since the document is enconverted into UNL, and has the combination of UNL sub graphs representing the conceptual features of the documents. The UNL based document specific features, and query document similarity score is computed to know the nature of the relevant document.

The following parameters are considered to find the document specific features.

- Occurrence of Multiwords (P_{MW}) within the document
- Occurrence of Named Entity(NE)(P_{NE}) within the document
- Total number of Relation Concept (CRC) (T_{CRC}/T_C) and maximum occurrence of CRC with tourism domain specific features(TourismCRC) within the document.

The Document specific feature score $D_s(q,d)$ helps to interpret the importance of the document with respect to the conceptual and linguistic features associated with the document.

$$D_s(q,d) = ((P_{MW} + P_{NE} + Tourism_{CRC}))/(T_{CRC} \vee T_C) \quad (4)$$

The document may be important because of the high frequent occurrence of Multiword or Named Entity or CRC/C that belongs to a tourism concept. The query document similarity score $Dquery_s(q,d)$ is computed based on the following parameters,

- Maximum occurrence of the query with the specific UNL relationship and it's UNL constraints. $p(r|q)$ is used to determine which relation r contributes more to a given query q in the document d. Similarly, $p(c|q)$ is used to determine which semantic constraint contributes more to a given query q in the document d.
- The Number of links for a query concept in a document(Number of queries CRC for that document)$(nLink_{Query}(q,D))$.

$$Dquery_s(q,d) = Argmax(p(r|q)) + Argmax(p(c|q)) + nLink_{Query}(q,D) \quad (5)$$

In order to normalize the ranking score between 0 and 1 the base ranking model is modified as follows.

$$LearnedRank((q_i, D_{ij})) = -(1log_2((q_i, d_{ij}) + D_s(q_i, d_{ij}) + D_{query_s}(q_i, d_{ij}))) \quad (6)$$

Hence, the document is ranked based on the $LearnedRank((qi, D_{ij}))$. A DCG score is computed in each iteration, and it is compared with that of the subsequent iterations. As discussed earlier, the learning iterations stopped when there is no change in the DCG score.

5 Result Evaluation

The objective of this learning to rank, is to find the learning capability to improve the relevance of the conceptual search results for different users. To this end, we evaluate the method by comparing the previous UNL based search and ranks, with the proposed learn-to-rank approach. We also compare the DCG value of per-user with different user feedback scores, and the overall ranking quality of our learning algorithm is measured by the nDCG score of the proposed approach with the existing methods. The training set consists of a different set of queries chosen from linguistic experts, and their associated results, which are ranked at different iterations, depending on the query and the user feedback. For some set of queries, the iterations stop at level two, and for some queries they go beyond two levels of iterations. We have taken 75 queries with different categories as described in section 3.1.1. The Tables 3 and 4 shows the normalized Discounted Cumulative Gain score for the base UNL search and Rank, and the comparison of a modified ranking model with per feedback and different user feedback.

Table 1. Comparison of nDCG score

Number of Methods	nDCG
UNL based search and Rank - Method 1	0.75
Learn to Rank with user feedback - method 2	0.86

Table 2. Comparison of nDCG-Single user with different user

Number of Methods	nDCG
Method 2 with single user feedback	0.71
Method 2 with different user feedback	0.86

The training corpus is tested for 10,000 documents from tourism corpus for 75 queries. The user feedback and documents conceptual features are learned for the top 20 retrieved documents. We found improvements in ranking different user feedbacks, when compared to per user feedback, and the proposed ranking model found 20% improvements over the base UNL based search and Rank models.

6 Conclusions and Future Work

In this paper, we have proposed a learningtoranking model, which changes the base ranking model by giving additional preference to the user feedback and its associated conceptual features. We have evaluated our result by comparing the results of the per user feedback with the different user feedback. We found good improvement in the different user feedback, compared to the per user feedback. We have also compared our base-level UNL search subsystem, and identified significant improvements in the DCG score. At present, though we consider the different user feedback, we have not utilized the personalized user search log. In future, we have planned to utilize personalized users search logs, and also include the page rank score, which can improve our ranking model in terms of a number of physical and conceptual links.

References

1. Li, H.: A short introduction to learning to rank. IEICE Transactions 94-D, 1854–1862 (2011)
2. Crammer, K., Singer, Y.: Pranking with ranking. In: NIPS, pp. 641–647 (2001)
3. Sellamanickam, S., Sundararajan, S., Garg, P., Selvaraj, S.K., Keerthi, S.S.: A pairwise ranking based approach to learning with positive and unlabeled examples. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM 2011, pp. 663–672. ACM, New York (2011)
4. Xia, Fen, Liu, Yan, T., Wang, Jue, Zhang, Wensheng, Li, Hang: Listwise approach to learning to rank: theory and algorithm. In: Proceedings of the 25th International Conference on Machine Learning, ICML 2008, pp. 1192–1199. ACM, New York (2008)

5. Radlinski, F., Joachims, T.: Active exploration for learning rankings from clickthrough data. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2007, pp. 570–579. ACM, New York (2007)
6. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: In Proceedings of SIGIR, pp. 154–161 (2005)
7. Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Chapelle, O., Weinberger, K.: Learning to rank with ($alotof$) word features. Inf. Retr. 13(3), 291–314 (2010)
8. (Undl foundation) Universal Networking Languages
9. Balaji, J., Geetha, T.V., Ranjani, Karky, M.: Article: Morpho-semantic features for rule-based tamil enconversion. International Journal of Computer Applications 26, 11–18 (2011), Published by Foundation of Computer Science, New York, USA
10. Weixin, T., Fuxi, Z.: Learning to rank using semantic features in document retrieval. In: Proceedings of the 2009 WRI Global Congress on Intelligent Systems - Volume 3. GCIS 2009, pp. 500–504. IEEE Computer Society, Washington, DC (2009)
11. Kuo, Wei, J., Cheng, Pu-Jen, Wang, Hsin-Min.: Learning to rank from bayesian decision inference. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 827–836. ACM, New York (2009)
12. Xu, Jingfang, Chen, Chuanliang, Xu, Gu, Li, Hang, Abib, Torres, E.R.: Improving quality of training data for learning to rank using click through data. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 2010, pp. 171–180. ACM, New York (2010)
13. Xue, Rong, G., Zeng, Jun, H., Chen, Zheng, Yu, Yong, Ma, Ying, W., Xi, WenSi, Fan, W.: Optimizing web search using web click through data. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM 2004, pp. 118–126. ACM, New York (2004)
14. Chen, N., Prasanna, V.K.: Learning to rank complex semantic relationships. Int. J. Semantic Web Inf. Syst. 8(4), 1–19 (2012)
15. Umamaheswari, E., Geetha, T.V., Parthasarathi, R., Karky, M.: A multilevel unl concept based searching and ranking. In: WEBIST 2011, pp. 282–289 (2011)
16. Subalalitha, T.V., Geetha, P.R., Karky, M.: Corex: A concept based semantic indexing technique. In: SWM 2008 (2008)
17. K, E., Geetha, T.V., Parthasarathi, R., Karky, M.: Core concept based query expansion, internet coimbatore tamil conference. In: Proceedings of World Classical Tamil Conference, Coimbatore (2010)

Tuning of Expansion Terms by PRF and WordNet Integrated Approach for AQE

Ramakrishna Kolikipogu and B. Padmaja Rani

Department of IT & Department of CSE
SWEC & JNTUH College of Engineering, Hyderabad, India
krikrishna.csit@gmail.com, padmaja_jntuh@yahoo.co.in

Abstract. Vocabulary mismatch in Information retrieval can be solved by Query Expansion (QE) techniques. Relevance feedback is a prominent solution to improve recall of retrieval system. Sometimes user may be reluctant and novice in providing feedback to improve the retrieval performance. Pseudo Relevance Feedback (PRF) automates the process. PRF treats top ranked resultant items are relevant and uses them to expand the query, which is not always correct. PRF by local analysis does not give guarantee to feedback positive terms to the system. Use of global analysis to capture the positive feedback is a regular practice in information retrieval process. This paper addresses the limitations of local analysis and global analysis alone by a novel approach that integrates both PRF and WordNet to select good expansion terms. The proposed solution filters the expansion terms and optimizes the expanded query. The proposed work is carried out on a huge Telugu text corpus collected from Wikipedia and other Telugu daily news portals.

Keywords: Query Expansion, Pseudo Relevance Feedback, Automatic Query Expansion, WordNet, Expansion Terms, Good terms, Information Retrieval, Indian languages, Telugu language.

1 Introduction

Availability of digital documents on internet and other information repositories are growing in an exponential way. No doubt that the search engines are helping end users to find and locate relevant items of interest expressed by supplying a query through interface. Use of Internet and other Information Accessing Systems are popular in Education, Medical, Business, Agriculture and other significant fields. Information Retrieval is a process of storing, retrieving and presenting the information to meet the interest of user queries. The result of the retrieval process depends on how well the IR system is designed to answer the following questions: 1) How well the Dataset is indexed? 2) How well the user writes a search query? 3) Efficiency of external resources used in processing the query and items? And of course the processor speed and network speed are the default issues to be considered for all the retrieval systems. From the last five decades the researcher found many methods to address the above queries

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 640–651, 2013.
© Springer International Publishing Switzerland 2013

and suggested different factors that they improve the search results. Indexing greatly influences the performance of retrieval system by depending on the levels of exhaustivity and specificity factors. Exhaustivity is the extent for which descriptors are included to cover maximum concepts. The exhaustivity is proportionally to the size of index. More the index descriptors occupy more size, almost double the database. There should be limited descriptors to represent an item of the dataset. Another factor specificity of indexing is the generic level at which the concepts assigned to the entity are expressed[1] An indexer may be unable to differentiate between specific descriptors, or information used in indexing may be insufficient for determining the most specific descriptor. These two factors measure the level of concepts covered by index descriptors. Low exhaustivity and high specificity or high exhaustivity and low specificity do not serve the purpose. Different Indexing techniques are being in use for information retrieval. Many indexing techniques are in practice which gives faster results. Automatic indexing techniques are proven to be effective in Information retrieval system. Statistical indexing was used in this paper to test the proposed work. The major problem in retrieval systems is writing of search query or topic by the end user. Terminology of nave user in writing a query is may not be strong and succinct to represent the concept of search and they terms are preferably simple and vague. Users prefer to input the short queries with one or two words, which may not resemble the interest of the user. In majority cases short queries fail to express the information needs of users to the system. Word mismatch and vocabulary mismatch are two severe problems which greatly affect the retrieval process. There is possibility of using poor vocabulary in expressing information needs, where as resources are drafted content with rich vocabulary that increases the vocabulary distance to the nave user query terminology, which is one of the reasons for IR system to perform in lower relevance levels. Query Expansion is a viable alternate to solve these problems. When user writes a query with insufficient set of terms the query expander generates a new query to expand the scope of the search. Relevance Feedback(RF)[2] is to improve the accuracy (Recall) by repeating the search with new query formed by user with feedback terms that are selected from known relevant items of initial search. The RF is an interactive feedback system that further increases overhead on users. This RF system is suggested for expert users, who can discriminate good and bad expansion features to improve their search results. It is not adequate for nave users, who cannot judge well and choose better relevant terms. Automatic Query Expansion(AQE) which also known as Pseudo Relevance Feedback(PRF) [3] techniques are viable alternate to RF techniques. Global analysis improves the scope to select more expansion terms from external resources like Thesaurus, WordNet, and Ontology etc. Too much use of external resources for expansion terms leads to query drift. This paper focus on how multiple resources can be used to optimize newer query which improves both recall and precision. We made an attempt to test and compare various Query Expansion methods with proposed system. The proposed model integrates local and global relevant sources to filter the expansion terms and see that the precision is improved without compromising recall.

2 Related Work

Abundant information resources created demand for IR systems. Commercial search engines are making huge population with the search users all over the world. No doubt Google, Bing, Yahoo, Ask, Aoi, and Mywebsearch are top 5 most popular search engines in terms of usage throughout the world. Digital information on internet is growing in multiple factor. Available Information in on-line consists of 20% numerical & 80% textual[4], where it is important to build effective text processing systems. Around 80% of Netigens search for information on internet and other repositories. Use of Local Languages for information retrieval is drastically increasing due to high availability of digital documents in those languages[5]. Effect of various methods used in IR may differ from language to language. It depends on the language features and their complexities. Word Mismatch is a common problem of Information Retrieval in every language. Query Expansion gives a solution to this problem[6, 7]. With the proliferation of the Internet in south Asia over the last decade, the availability of digital documents in Indian languages has been increased considerably. The need for effective information access methods for these languages is being increasingly felt[8]. Primary typing mistakes and spelling errors for telugu are addressed by G.U.Rao in his recent work[9], a spell checker. Popular techniques had shown a severe problem of *low recall* while accessing information using Telugu and other Indian language queries[10]. Due to morphological richness, Telugu Information Retrieval systems severely suffering from word mismatch, this is not only with TELUGU, but also for many other languages.

2.1 Query Reformulation

Amanda Spink, Dietmar Wolfram and other[11] analyzed that the average length of a user query on the web search is 2.4 terms. They further analyzed the behavior of search users that a user always preferred to enter short query and expects the relevant outcome in the top 10 pages. However, queries can fail to find documents of interest due to a mismatch in terminology. AQE supplements this problem by adding new terms from top ranked items and formulate a new query to repeat the process. While query expansion has been shown to be effective at improving query performance in terms of recall, will slower the retrieval process. Vocabulary mismatch is one of the major causes of poor recall in Information Retrieval. Resource builders and users invariably select different set of lexical units to specify their interest of search, causing retrieval methods based on lexical matching to miss relevant items[12]. Expansion of query at search time is called run time query expansion. The context of selecting expansion terms is classified in to Local and Global. Local analysis is one of the successful techniques of query expansion models[13]. In local analysis a set of initial resultant items are analyzed to expand the search either by user or system itself. One of the well known interactive local method is RF[14], which requires a user to repeatedly interact and judge the relevance of output items. RF is used to expand the boundary of a query using judgment, when items retrieved by initial query

are relevant. The J.J.Rocchio in 1971 through SMART[2] retrieval system proven that, along with adding relevant term weights to the query and negating non-relevant term from the query will further improve the recall of search. The main objective of RF system is to improve the relevance of search through multiple iterations adjudged by end user. Another local analysis policy called Automatic Relevance Feedback(ARF) aims to reduce overhead of the user and refine the results. ARF is to avoid user interaction to provide relevant expansion terms to the system, ARF or PRF system automates the judgment and focus on top resultant items to improve the results. User feedback can considerably increase effectiveness[15], but further increase overhead on user and misleading judgment by non expert users totally miss the context of search, the automatic expansion is more likely to lead to better performance in this regards[16]. There are many factors to be considered while implementing AQE. Before expanding the query, machine training is required. AQE or Blind Relevance Feedback (BRF) has a long history in information retrieval[17], as it has been notified before 1960s[18]. In PRF the query is added with more terms from top-ranked items to expand the query[19]. PRF technique is an effective approach to improve the recall and increase the effectiveness of the search[19–23]. Augmenting of expansion terms, cause query drift[24]. The deviation of root concept can be from selection of non-related terms from the results. The terms correlated with each individual term of the root query rather than with the entire query probably match unrelated terms. This scenario is more dangerous, if the selected term is a proper noun[25]. If the entire expansion terms collection is not properly selected inline to query concepts based on initial top resultant items, this may cause decrease in precision. Instead of simply selecting expansion terms from top ranked items of initial query, many researchers adopted new factors in choosing expansion terms and controlled fall of precision. Xavier Tennier and Clement d Groc [26] modeled term co-occurrence in a fixed window with random walk algorithm to select expansion terms through PRF approach not exaggerated the recall , but it helps to control the loss of precision. As PRF is a two phase retrieval process, it consumes time and delays the final results. Hao Wa and Hui Fing[27] proposed an incremental approach which reduces the time to reformulated the query by adjusting the scores in the document accumulators as expander terms. With the due study, PRF techniques are suffering from query drift, that the reformulated queries miss the context of initial search. Time taken to return the search results is one of the performance aspect, but not more important than the relevance of the search. Relevance of the search can be preserved by holding query concept in expansion process. As the concepts are more important to preserve the context of search query, the expansion terms are selected from concept resources such as knowledge structures like Thesaurus, WordNet or Ontologies etc. Use of semantic networks to expand the query is a better idea which will improves the recall and precision. WordNet is an online lexical database that provides different levels of concepts to different vocabulary terms. Coverage of WordNet influences the expansion terms selection in retrieval system. The distance between terms may broaden or narrow down the search. Voorhees[28] used semantic relationships

to expand query by using WordNet database The task is found to be difficult for specific ad hoc retrieval, that the knowledge base is designed for general purpose. AQE by connecting to a knowledge structure WordNet is successfully done by Shuang Liu, Fang Lui et al[29]. They used WordNet to disambiguate the sense of a phrase and added the supplements to the query. They had shown that the results by this method gain 23% to 33% over the best known results on TREC 2009, 2010 and 2012 track collections for short queries without using web data. The impact of both local analysis and global analysis for query expansion is compared by Jinxi Xu and W Bruce Croft[3]. They concluded that the local analysis is quite better than global analysis. Use of WordNet and ConceptNet together improve the retrieval performance[29]. Zhiguo Gong, Chan Wa Cheang, U Leong Hou[30] found that the use of WordNet can improve the results, but the efficiency of WordNet is a big problem for which they proposed an alternate semantic structure called Term Semantic Network (TSN). TSN included term co-occurrence and semantic relatedness to reduce the noise rate of WordNet. We found that use of query expansion solves the word mismatch and vocabulary mismatch problems in retrieval process. Local analysis is more effective than global analysis. Relevance feedback to reformulate the query after user judgment improves recall, but user need relevant knowledge to judge the initial results. Domain search by expert user may satisfy more with relevance feedback. Automatic relevance judgment by machine with proper training will improve recall and precision. The way how expansion terms are selected is more important in query expansion , wrong selection causes outlier retrieval.

3 Selection of Good Expansion Term

Selection of expansion terms might be as good as possible to optimize the search and improve relevance. In this paper, we proposed a novel approach to select good expansion terms in two levels. In first level, expansion terms are being selected using local analysis with Pseudo relevance Feedback. In second level, expansion terms are chosen from WordNet by global analysis.

3.1 Pseudo Relevance Feedback

PRF method considers that the top ranked items are relevant to the initial query. PRF automatically adds top terms to the initial query and creates a new query to repeat the search process. PRF will reduce overhead of the user and broaden the scope of iterative search. We consider a user query q is a vector $\boldsymbol{q}$ and $\boldsymbol{q} = (q_1, q_2, q_3, ...q_n)$, where each q_i is a weight of query term contained in query q. Query term weights are calculated by SMART system[31]. Documents are indexed by Inverted index with Vector Space Model(VSM) and each document in the database is further represented as document vector $\boldsymbol{d} = (d_1, d_2, d_3, ...d_m)$, where each d_j represents a document term weight. Document dataset D consists of collection of items. In this work we represent documents and queries as weighted vectors using term frequency tf_{ij} and Inverse Document Frequency idf_i. Throughout the work, construction of vectors follow tf and idf measures.

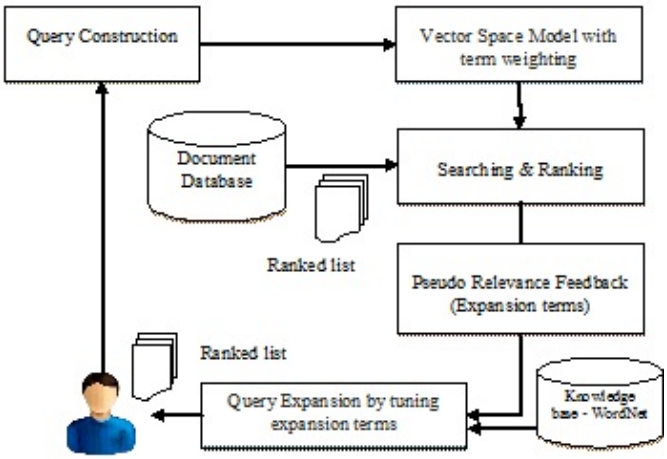


Fig. 1. Selection of good expansion terms by filtering PRF and WordNet based expansion terms

$$idf_i = \log(\frac{N}{df_i}) \tag{1}$$

Where N is total number of document in the database and df_i is document frequency over the collection which contain term i.The weight of each tem in a document is measured by dot product of tf_{ij} and idf_i.

$$d_{ij} = tf_{ij} * idf_i \tag{2}$$

Where , idf_i is inverse document frequency, which measure the importance of a term throughout the collection. User query is weighted in the same way how document is weighted using function (1) and (2). The work is tested on a Telugu language text corpus containing 124000 items. WX-Notation (Romanization) is used for pre-processing and post-processing the Telugu text. WX-Notations for Telugu with WX2UTF and UTF2WX converters[32][33] are used throughout the work.

For the query q_1 "kriketar jattunu AxukunnADu "which means "Cricketer supported team "is represented as query vector."kriketar/cricketer, jattunu/team, AxukunnADu/supported ". After initial search it returned the following 3-documents as top results.
Doc-1 : *mAji ICL wripura kriketar mahixivAkar wripura jattunu AxukunnADu*
Doc-2 : *ICL kriketar rashmIxAs jArkhAmD ICL jattunu AxukunnADu*
Doc-3 : *kricketar rashmIxAs vikeT nu paDagoTTi kriketar jahIr khAn jattunu AxukunnADu*
The expansion terms from these top 3 items are selected using sorted list of ranked term from resultant item set based on the occurrences f of term t in the result collection.

Table 1. Ranked list of relevance terms based on f or df is given in the table for top-3 ranked items

#S.no	term-t	tf	df	f	idf	f*idf
1	mAjI	1	1	3.18	3.18	3.18
2	ICL	2	3	2.88	5.76	8.64
3	kriketar	3	4	2.70	8.1	10.8
4	mahixivakar	1	1	3.18	3.18	3.18
5	xripura	1	2	3.18	3.18	6.36
6	jattu	3	3	2.70	8.1	8.1
7	AwukunnADu	3	3	2.70	8.1	8.1
8	rashmixAs	2	2	2.88	5.76	5.76
9	jArkhamD	1	1	3.18	3.18	3.18
10	vikeT	1	1	3.18	3.18	3.18
11	paDagoTTi	1	1	3.18	3.18	3.18

The top n terms are selected as expansion term. In this example if we consider top-5 ranked term are fig8.jpg according to their f factor measure.

Table 2. Ranked expansion terms using Pseudo relevance Feedback approach

#S.no	term-t	df	f	idf	f*idf
1	kriketar	3	4	2.7	10.8
2	ICL	2	3	2.9	8.64
3	jattu	3	3	2.7	8.1
4	AxukunnADu	3	3	2.7	8.1
5	xripura	1	2	3.2	6.36

At the end of PRF, we selected a set of expansion terms and represented them in to vectors as query vector Q_{PRF}.

3.2 Selection of Expansion Terms from WordNet

WordNet ontology is one of the most important lexical data resources used in the field of text analysis, computational linguistics, and many other information retrieval related research areas. WordNet[2] is ontology of lexical references whose design was inspired by the current theories of human linguistic memory. Nouns, verbs, adjectives and adverbs are grouped into sets of synonyms called synsets, each representing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations such as hypernym/hyponym (IS_A), and meronym/holonym (PART and WHOLE). Query Expansion based on WordNet synonym relations will improve the recall and precision. Use of synonyms, different concept with same sense doesnt change the context of search. S_1 and S_2 are synonyms, if they can interchangeably use with same sense in same context. In this work retrieval had considered synonyms of Synset as expansion terms to

improve the accuracy of the search. Each query term have a set of synonyms in a lexical database. In this work we considered only synonyms of the query terms as expansion concepts, due to lack of complete WordNet for Telugu. A synset of training corpus terms collection is manually created and used in proposed system. On an average each term (Nouns, Verbs and Adjectives) of 20% of word corpus is taken and prepared synonyms for each possible words. On an average 2.6 synonyms for each word are created manually and stored them in to corresponding Synsets.
1) $kriketar \rightarrow \{kriketar, AtagADu\}$.
2) $jattu \rightarrow$ {bRmxam, bRnxam, cayam, Dambaram, gaNam, gumpu, gumi, jAlam, jattu, kUDika, kUtami, kuxuva, kakshahi, kalApam, kurumbam, lATu, lompu, mamDali, mogi, mollam, mollami, mowwam, mutAnicayam}.
AxukunnADu/helped is a variant form of root word "*AxukO* "/ "*help* ". The root forms along with tagging of inflated variants are extracted by using morphological analyzer[34].
3) *AxukO* →{(cEyUxanivvu, sAya-paDu, vennu_xannu, xODpATu-paDu, xODpaDu, xODupadu, sAyambaTTu, sAyaxApaDu}.

The set of expansion term are represented as WordNet(WN) based query vector Q_{WN}. The weights of all synonyms of a query term are equal. Synonyms carry same meaning without changing the context and sense of use. These synonyms are chosen as expansion terms from WordNet (i.e Synset for this study). Many researchers failed to improve retrieval performance by using WordNet. The synonyms may carry same meaning, but if those words cannot appear in any of the documents will increase the search time and fails the search at the end. To overcome such out-of-boundary problem, we proposed a novel method to filter the expansion terms.

3.3 Filter and Tune the Expansion Terms to Reformulate the Query

Instead of directly using expansion terms selected from PRF and WordNet in reformulating query, our idea is to clean the expansion terms, which selects appropriate set of expansion term to improve the recall without compromising precision of retrieval system. Expansion terms Q_{PRF} & Q_{WN} are used to define final set of expansion terms.

$$Q_E = Q_{PRF} \cup Q_{WN} \tag{3}$$

Where Q_E, is filtered and Expanded Query, Q_{PRF} is PRF based expansion terms and Q_{WN} is WordNet based expansion terms. Each expansion term added to new query list from PRF and WordNet are assigned new weights. Before considering expansion terms to the final set, a look-up list is verified for possible existance of such a term in any of the items. Once the occurance of expansion term is confirmed with dictionary look-up table, the weights are caluculated and each synonym is given eaqual weight.

$$\cos\theta = \frac{(d_j \bullet q)}{(|d_j| \bullet |q|)} \tag{4}$$

Where $d_j = (w_{1,j}, w_{2,j}, w_{3,j}...w_{n,j})$, $q = (w_{1,q}, w_{2,q}, w_{3,q}...w_{n,q})$ are document vector and query vector respectively. The another similarity coefficient between query q and document d can be usually expressed as $sim(q, d)$, dot product of query vector and document vector.

$$sim(q, d) = \sum w_{t,q} \bullet w_{t,d} \tag{5}$$

Where t is same term from query q and document d. The similarity scores are calculated for each document with respect to expanded query and ranked the order of relevant items in decreasing order.

4 Results Analysis

The entire work had been implemented for Telugu language over a set of 124000 text documents with 50 queries. Entire dataset is properly indexed in inverted file structure as weighted vectors. 20 single term queries were tested on proposed model along with base models. 30 Multi-term queries were also tested and results were plotted below. Average length of queries used in the proposed system was 3 terms. Below figure shown the effect of Pseudo relevance feedback over normal query search. The performance of retrieval system is good comparing with baseline approach, but precision remains stable between 0.6 and 0.7, 0.9 and 1.0 recall points. The performance of retrieval system using WordNet is shown in Fig. 3. The average relevant documents are almost null after retrieving the last relevant item in either baseline search or expanded search using synonyms.

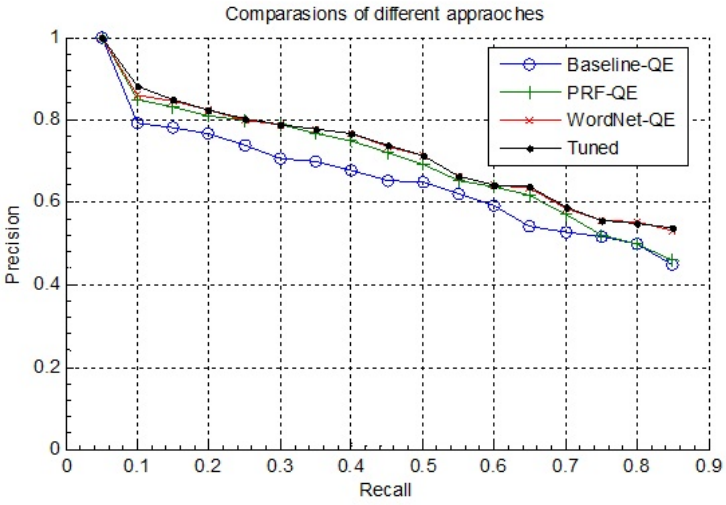


Fig. 2. Recall-precision comparison between Baseline retrieval without QE and QE with PRF,Synset,Proposed approach

5 Conclusion

Word mismatch and Vocabulary mismatch are two severe problems information retrieval. Query Expansion supplements the problem and improves the search scope with and without user intervention. Relevance feedback is proven to be effective local analysis based query expansion technique, but it require user interaction to feedback the relevance of search. Users may be reluctant or novice to feedback relevance to repeat the search. Automatic feedback is a viable alternate to relevance feedback approach. It blindly supplies expansion term by considering top ranked item as relevant from initial search process. It may adverse the purpose of search and miss the context. Too much of using expansion terms lead query drift. Semantic network and knowledge bases provide good support to expand the search without missing initial context of the search. Generalization and specialization of expansion terms retrieve more unrelated documents and further decrease the relevance. Controlled language resources like Synset a part of WordNet effectively support query expansion, but it takes more time when the supplemented terms (synonyms are more). Selection of good expansion terms is a challenging task for researcher. Co-occurrences, ontology structures, query logs, link analysis, personalized expansion, behavior analysis are currently focused by many research to improve the search results. We proposed a new method to find better alternate terms to expand the query. Our method extracts initial expansion terms from Pseudo Relevance Feedback and filter them by connecting to a knowledge structure WordNet and tune to optimize expansion terms list, which improve the recall without loss of precision. The results had shown that the recall and precision were increased when compared to baseline search, PRF and WordNet based search. We planned to continue this work by considering new factors to expand the query. This idea can be used in cross-lingual information retrieval as knowledge resources are available for many languages.

References

1. Soergel, D.: Indexing and retrieval performance: The logical evidence. Journal of the American Society for Information Science (May 1994)
2. Rocchio, J.J.: Relevance Feedback in Information Retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing (1971)
3. Xu, J., Bruce Croft, W.: Query expansion using local and global document analysis
4. Conlon, S., Lukose, S., J.G.H.: Automatically extracting and tagging business information for e-business systems using linguistic analysis (2008)
5. Hoeber, O., Yang, X.-D., Yao, Y.: Conceptual query expansion. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) AWIC 2005. LNCS (LNAI), vol. 3528, pp. 190–196. Springer, Heidelberg (2005)
6. Egidio Terra, C.L.A.C.: Scoring missing terms in information retrieval tasks. In: Proceedings of International Conference on Information and Knowledge Management - CIKM, pp. 50–58 (2004)
7. Wei, C.-P., Hu, P.J.-H., C.H.T.: Managing word mismatch problems in information retrieval: A topic-based query expansion approach. Journal of Management Information Systems 24(3), 269–295 (2007)

8. Prasenjit Mujumder, M.M.: Indian Language Information Retrieval. In: Guide to OCR for Indic Scripts Advances in Pattern Recognition (2010)
9. Patel, D., Madalli, D.P.: Scoring missing terms in information retrieval tasks. In: Proceedings of International Conference on Information and Knowledge Management - CIKM, pp. 50–58 (2004)
10. Prasad, P.V.: Recall Oriented Approaches for improved Indian Language Information Access. In: International Institute of Information Technology, Hyderabad, India (2009)
11. Spink, A., Wolfram, D.,M.B.J.J.: Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology 52(3), 226–234
12. Salton, B.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science 41(4), 288–297 (1990)
13. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems 18(1), 79–112 (2000)
14. Christopher, D., Manning, H.S., Raghavan, P.: An Introduction to Information Retrieval. Cambridge University Press (2009)
15. Magennis, M., van Rijsbergen, C.J.: The potential and actual effectiveness of interactive query expansion. In: Proceedings of ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval, New York, pp. 324–332 (1997)
16. Ruthven, I.: Re-examining the potential effectiveness of interactive query expansion. In: Proceedings of ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval, New York, pp. 213–220 (2003)
17. Carpineto: Automatic query expansion in information retrieval. ACM Computing Surveys (CSUR) 44, 1–49 (2012)
18. Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. Journal of Association of Computing machinery 7(3), 216–244 (1960)
19. Robertson, S.E., Walker, S.: Okapi or keenbow. In: Proceeding of Text Retrieval Conference (TREC), pp. 151–161. NIST Special Publication, Gaithersburg (1999)
20. Glen Robertson, X.G.: Improving abraq: An automatic query expansion algorithm. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (August 2010)
21. Ian Ruthven, M.L.: A survey on the use of relevance feedback for information access systems. The Knowledge Engineering Review 18(2), 95–145 (2003)
22. Le Zhao, J.C.: Automatic term mismatch diagnosis for selective query expansion. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 515–524 (August 2012)
23. Ramakrishna Kolikipogu, P.R.B.: Wordnet based terms selection for pseudo relevance feedback model. In: Proceedings of 3rd International IEEE Conferences on Computing, Modeling and Simulation, Mumbai, India, January 2011, pp. 127–131 (2011)
24. Mitra, M., S.A., Buckley, C.: Improving automatic query expansion. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 24-28, pp. 206–214 (1998)
25. Vechtomova, O., Karamuftuoglu, M.: Elicitation and use of relevance feedback information. Information Processing Magazine 42(1), 191–206
26. Tennier, X., de Groc, C.: Experiments on pseudo relevance feedback using graph random walks. In: Proceedings of String Processing and Information Retrieval, Canada de indias, Colombia, October 21-25, pp. 193–198 (2012)

27. Hao Wu, H.F.: An incremental approach to efficient pseudo-relevance feedback. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, August 1, pp. 553–562 (2013)
28. Ramakrishna Kolikipogu, P.R.B.: Reformulation of telugu web query using word semantic relationships. In: Proceedings of ACM-International Conference on Advances in Computing, Communications and Informatics, Chenai, India, pp. 774–780 (October 2012)
29. Liu, S, L.F.Y.C.T.M.W.: An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: Proceedings of the 16th ACM-International Conference on Database and Expert Systems Applications, Sheffield, UK, July 25-29, pp. 266–272 (2004)
30. Gong, Z., Cheang, C.W., U.L.H.: Web query expansion by wordnet. In: Proceedings of the 16th ACM-International Conference on Database and Expert Systems Applications, Copenhagen, Denmark, August 22, pp. 166–175 (2005)
31. Salton, G.: The SMART Retrieval System Experiments in Automatic Document Processing. Prentice-Hall, Upper Saddle River (1971)
32. Rohit Gupta, P.G., sapan: Transliteration among indian languages using wx notation. In: Proceedings of Konvens 2010, Saarbrcken, Germany, pp. 147–150 (September 2010)
33. Ramakrishna Kolikipogu, P.R.B.: Study of indexing techniques to improve the performance of information retrieval in telugu language. International Journal of Emerging Technology and Advanced Engineering 3(1), 482–491 (2013)
34. Uma Maheswara, G., Amba kulkarni, C.M.: A telugu morphological analyzer. In: Proceedings of International Telugu Internet Conference, Milpitas, US, pp. 1–7 (February 2011)

Concept Based Personalized Search and Collaborative Search Using Modified HITS Algorithm

G. Pavai, E. Umamaheswari, and T.V. Geetha

Anna University, Chennai, TamiNadu, India
pavaigops@yahoo.co.in, umavasanth28@gmail.com, tvg@hotmail.com

Abstract. Keyword based search is commonly used by popular search engines. The major problem with this kind of search is that we do not get user intended results for the search. In addition, every user gets the same set of results for the same query whereas, their interests may be different. In order to tackle this, we go in for personalized web search and collaborative web search. We find out the user interest and accordingly display only pages that are relevant to their interest and not relevant blindly only to their query. This paper, describes a novel approach for storing the personalized user concepts and proposes a modification to the HITS algorithm based on user interested concepts. This paper also describes how to extend the concept based personalized search to concept based collaborative search. In addition we propose a new methodology to form dynamic groups in the case of collaborative search.

Keywords: Concept based search, Personalization, Collaboration, Concept based HITS, User profile.

1 Introduction

The amount of information on the web is rapidly growing. Information retrieval systems are responsible to provide the information of interest to users of the systems. Users typically type a query consisting of a few keywords describing what they need. Information Retrieval systems perform a word to word match of the query words with all the documents in their collection and return matching documents. Web Retrieval is much more complicated due to the large and dynamic content on the web. The web search engines usually serve millions of users and millions queries every day. It is very unlikely that all of those users have similar interests and search for similar information.

However, users use query words that have ambiguous meanings like 'play' used to mean play-play some game or play- the drama. Therefore, for different user backgrounds, different user interests and ambiguities in natural language, it is very likely that query words of two different users may be exactly same even though what they need is different. This is seen clearly in the following three cases: (1) When a query contains ambiguous keywords. Different users may use exactly the same query (e.g., play) to search for different information but existing IR systems return the same results for these users. Without considering the actual user interest, it is impossible to know which sense "play" refers to in a query. (2) When a query contains partial information: A query may contain an acronym or a shorter usage of a longer phrase. Then there

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 652–662, 2013.
© Springer International Publishing Switzerland 2013

might not be sufficient information required to infer the information need of the user. For example a query like ”IR” can mean ”Indian Railways” or ”Information Retrieval” among others. Existing systems return mixture of results containing the exact word which might contain different expansions. Knowledge of user interests and/or location of the user could be helpful in gathering more information required to understand what the query is about. (3) When information need of the user changes: A users information needs may change over time. The same user may use ”IR” sometimes to mean the Information Retrieval and some other times to mean the Indian Railways. Without recognizing the search context, it is impossible to recognize the correct user sense. Thus using user context information about the user and the query is necessary for improving the search accuracy. A more appropriate query that better explains the scenario is ”bus services in Java”. Does it mean Java - the programming language or Java - a place.

This is done using the techniques of personalized and collaborative web search where the user interest is learnt through the past history of the users and ranking of pages for the current query is based on previous interests of the user. Some amount of work has already been done towards the improvement of the performance of the search engine by storing user interested keywords which is described in the related work section. This is followed by the description of our work which in turn is followed by the results and evaluation.

2 Related Work

From when search engines came into existence, various methods were proposed for search engine enhancement. Personalizing web search results is essentially tuning future search results to be shown to a person based on his past, i.e. a person interested in a page earlier is likely to visit similar pages in the near future. This is one of the hot areas of research today. Susan et al., [9] gives a detailed classification of how these user profiles are learnt and represented in a personalized system. The classification is based on where the process is applied (client or server), how data is collected (implicit/explicit), where from the data is collected (browser cache, web server logs, proxy servers or browser agents), how user profiles are represented (weighted keywords, weighted concepts, or semantic networks). While Susan et al., [9] only survey the user profile component of personalized search, we do have another side of personalization where classification can be based on where in the search process the modification is incorporated - in indexing, searching , query processing, ranking or re-ranking. However, our profile representation is different from the ones described in this paper. We have used the multi-list structure for profile representation. Apart from the contribution to the user profile construction, we have incorporated the modified HITS algorithm in the ranking module to support personalization and collaboration in search engines.

Sendhilkumar and Geetha, [8] use a specially designed browser, for collecting user information, and uses ODP taxonomy to give categorical labels to user interests. However, using this system of profile creation becomes a question mark without this browser. Instead of categorical labels, we have used the UNL (Universal Networking Language) representation where we have assigned the concept to each word. Jia and philip [2] uses bookmark information to create user profile based on the term spans, term specificity

and image term extraction. However it was found that bookmarks that were identified were not very relevant to user interests. So, we are not using them in our work. Pallavi et al., [6] tackled this problem by using not only the browsing history of the users but also, the correlation between the query terms and the document terms selected by the user. But, this method does not give good results for specific queries and is not able to handle frequent changes in user interests. The security issue of Personalization is handled effectively by creating dynamic user profiles and ontology to represent user interests Namita et al., [5] and collaboration for handling frequent change of user interests. Yang and Wang [12] handled the same issue differently by using personalized files that are sent to the server along with the search query. In this work, crawling is done online based on the query and the file is sent along with the query. This reduces the job done by the server since the crawl itself is personalized. Whereas, Hochul et al., [15] tries to correct the security problem by using collaboration i.e. a user is grouped into a group of other users who have similar search behaviors.

Personalization's important variation that is widely explored is collaboration which is further subdivided into Asynchronous (different time different place) and Synchronous (same time different place). Delgado et al.,[1] deals with Asynchronous collaboration, where along with the correlation between the users, similarity between the topics chosen by the users are also considered. This improves the similarity of a user with a group and the user is likely to get more relevant results. Jose et al.,[3] also deals with Asynchronous collaboration where collaboration is clubbed with content based retrieval to further enhance the search results. Roman et al., [7] also describe Asynchronous collaboration where user centeredness is not only based on the particular user but also on how he forms part of the community. Susan et al., [18] Sendhilkumar and Geetha [19], Mariam et al., [20], Kenneth et al., [21] and Prabaharan et al [22] have described concept based personalization using domain ontology. The basic difference between their work and ours is in the profile creation. For example, Sendhilkumar and Geetha [19] have used the Personalised Page View (PPV) graphs for profile creation where as we have used the multi-list structure for profile creation. They have used ontology for semantics whereas, we have provided semantics using the Universal Networking Language (UNL).

Personalization's other variations include demographic personalization where demographic information about the active user is considered in order to identify neighborhood based on demographic similarity is discussed in Hu et al. [13], knowledge based recommender systems described by Shi, [14] where a knowledge model of a domain is used to retrieve pages based on the features they have and how these features derive to the user needs, content based filtering described by Hochul et al., [15] uses personalizing the search based on the content of the web pages in the form of features.

While the above works are descriptions of systems as a whole, we also have the ranking algorithms that enhance the existing systems. The Widely used HITS algorithm, uses hubs (pages with many outlinks) to calculate the authorities (pages with many inlinks). The authorities are nothing but the more popular pages. The problems with the traditional HITS algorithm Kleinberg [4], as per Joel et al., [16] are non - uniqueness(convergence of authorities may not be unique) and nil - weighting(a zero authority score for some of the authorities of the subgraphs). Joel et al., [16] proposes three new

ideas to handle these problems. The first one is using an exponential valued adjacency matrix instead of the traditional binary valued approach, the second one is using usage weighted input (the usage weights are obtained from the web server logs), and the third a combination of the above said two methods. J.Jayanthi et al., [17] shows another variation of the HITS that also uses Rocchio for feedback. This method uses two matrices called the Document- Category(DC) and the Document- term(DT) matrices. Every user is mapped to a category where weights are assigned based on the search history. When a user asks a query, the links in both the DT and DC matrices are prioritized. While each of the above methods modified the HITS algorithm from different perspectives, we have given a concept based modification to it to support personalization and collaboration in search engines.

3 Methodology

We describe the existing methodology in section 3.1 followed by the description of our work in section 3.2 and section 3.3.

3.1 Existing Methodology - Baseline

COREX system of ranking proposed by E. Umamaheswari et al., [10] is followed by the COREE Search Engine which is a Concept based search engine for the Tamil language. We use this system as the baseline for the evaluation of our work.

The UNL [11] representation is in the form of semantic networks with hyper-nodes where nodes represent concepts and arcs represent relations between concepts. Concepts are annotated.The search engine uses the UNL framework for representing the concepts. Every document considered for the search process in the COREE Framework is indexed in the form of three trees namely the C Tree (for concepts), the CR Tree (for Concept - Relations) and the CRC Tree (for the Concept - Relation - Concepts) and the process is clearly described by E. Umamaheswari et al.,[10]. Fig. 1 gives a detailed account on how to find the C's, CR's and CRC's of a sentence.

All the documents are indexed in all or any of the above categories based on their content. A three level ranking methodology is followed here.

Level 1: Degree of match categorization prioritizing documents based on complete match (CRC match), partial CR match or Concept only(C) match.

Level 2: Concept association categorization that is based on whether the match is a term match, concept match or an expanded concept match.

Level 3: Ranking is based on index based features like the frequency of occurrence of the term and concept in the document, position weight, Named Entity (NE) weight and Multiword (MW) weight.

3.2 Concept Based Personalization

This modified version of the HITS algorithm is called the PNHITS or the Personalised HITS. Figure. 2 gives the personalization framework and it is described in detail in the following sub-sections.

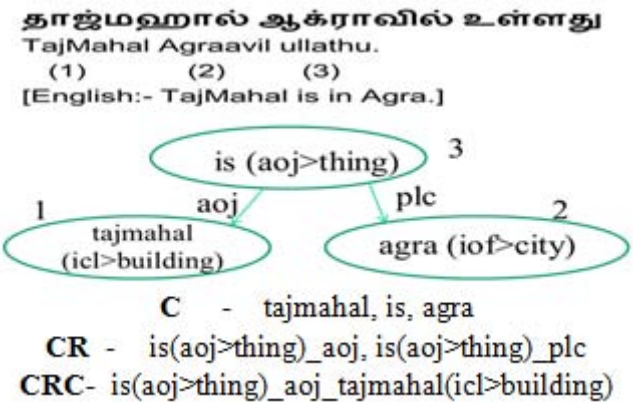


Fig. 1. Example for C, CR, CRC in a sentence

User Interest Tracking. A user's time spent on a web page is tracked via session tracking in JSP. If the time spent on a web page is higher than the initial threshold of two minutes, then the page is classified as interesting for the user for the current query. When a user asks a query, the stored user interested concepts are used to rank the web pages in a different manner such that the results obtained are satisfying the user's interest.

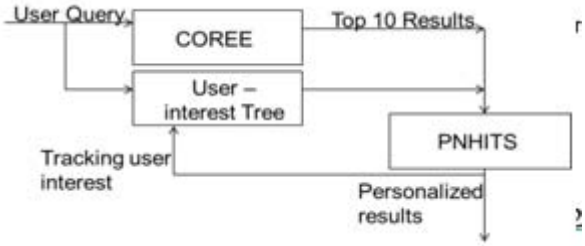


Fig. 2. Personalization Framework

User Interest Representation. The user interest is represented in the form of a Multi list data structure as given in Figure. 3 where each of the fields are described as follows.

- ASCII value for the username- for comparison purposes.
- User name- name of the user.
- Query concepts- the query concept which the user is interested in.
- Interested concepts- the concepts that a user is interested in for the corresponding query concept.
- A link to the left sub-tree.

User name Identifier	User name	User's Interested query	Interested concepts for the query	Link-to-left Sub tree	Link-to-Right Sub tree	Link to next interesting query concept

Fig. 3. USER - INTEREST Tree

- A link to the right sub-tree.
- A link to the next interesting query for the user.

A link to the next query- If there are more than one query concept for the user, we represent them by using a linked list. The pages visited by a user for user-interested queries are tracked to find out the user interested concepts for each user and are stored in a multi-list tree structure as shown above, for easier retrieval. In future, if the user asks a query, the tree is searched to find if the query has any concept similar to the interesting query concepts stored for that user. If a match is found, we prioritize the ranks of pages with user interested concepts using the PNHITS algorithm.

Search Process. When a new query is given, the COREX methodology is followed for ranking the web pages. The first ten ranked documents are given as an input to the PN-HITS ranking algorithm. These ten documents form the root set. The base set is formed by expanding these ten document's links. This results in a base set with a substantial amount of documents. A run-time adjacency matrix is formed for the documents in the base set depicting their link structure. Now, the current user's profile is searched for a match for the current query context. If a match in the tree for query is found, the interested concepts pertaining to the query are retrieved from the multi-list tree for using in PNHITS and pages represented in the adjacency matrix are ranked higher if they have the retrieved user interested concepts. If the retrieved page has a matching C, then it is multiplied by a factor of 1, a matching CR has its link value by 1.5 and tit is 2 for a matching CRC.

Given below is the description of the HITS algorithm [4]. Given weights $x^{<p>}$, $y^{<p>}$ the I operation updates the x-weights as follows.

$$X^{<p>} = \sum Y^{<q>}$$

$$q : (q,p)eE \tag{1}$$

The O operation updates the y-weights as follows.

$$Y^{<p>} = \sum X^{<q>}$$

$$q : (p,q)eE \tag{2}$$

In our work, the HITS algorithm is modified by introducing changes in the calculation of the hubs and authorities. As, a result, the equations 1 and 2 are changed as follows:

I operation:

$$x^{<p>} = \sum(y^{<q>} + cw * fac)/2$$
$$q : (q,p)eE \tag{3}$$

O operation:

$$Y^{<p>} = \sum(X^{<q>} + cw * fac)/2$$
$$q : (p,q)eE \tag{4}$$

Equations 3 and 4 represent the modified version of the traditional HITS algorithm, where
cw = 1, if the web page has the user interested C in its page.
cw=1.5, if the web page has the user interested CR in its page.
cw=2, if the web page has the user interested CRC in its page.

The concept weight (cw) is multiplied by a factor called fac. This is done in order to rank pages with the CRC higher than pages with CR relation and the pages with C relation are ranked lower than the pages with the CR relation. The weights for cw given above are based on the heuristic that a document with the CRC in the query is more important than the document having only a CR or C of the query as per E. Umamaheswari et al., [10]. Through experiments, we found that if the value of fac is 2.2, there is a good improvement in the ranks. But in general, any value that is greater than one would suffice. Ranking the pages based on their descending order of the authority score will give us the results that are personalized for each user. The PNHITS ranking algorithm boosts up the ranks of the direct out-links of the pages that are personalized along with the boosting of the ranks of pages that have the relevant personalized concepts. In short, the traditional HITS algorithm ranks the pages with greater number of in-links higher in the order, whether or not the page contains personalized user concepts. Whereas, the PNHITS algorithm ranks pages with higher number of personalized concepts and at the same time having a higher number of in-links higher than the other pages thereby making the highly popular pages (with more in-links) containing the user interested concepts appear first in the ranked order than the other pages.

3.3 Concept Based Collaboration

The user interest tracking, the user interest representation, and the ranking algorithm are the same as that of Personalization module. But, the retrieval process is different. Even, if the query has no matching user interested concept for the current user, it searches for the query concept as the user interested concept of any other user and if a match is found the rank of those pages are boosted.

The Search Process. For the ranking purpose we use the PNHITS algorithm described earlier. Initially, we ranked pages for collaborative search based on considering the

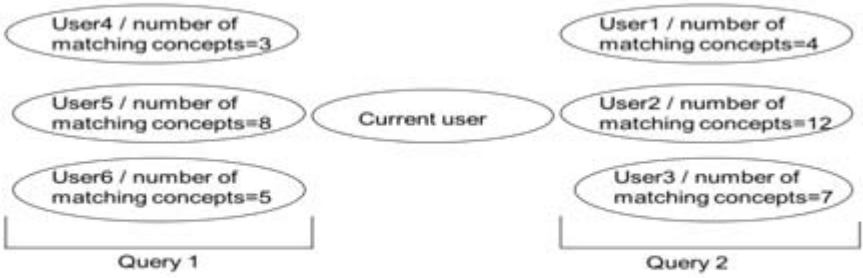


Fig. 4. Group Formation-scenario for two different queries

whole lot of users as one large group. But this method has the disadvantage of considering one user's interest as relevant to all other users. But this is not true. So, we go in for multiple groups.

The new search strategy is as follows: When a user gives a query, we check if there is a match for the same query for the current user and if any match is found, we retrieve those results. In case no match is found, we check for the same query concept as the interested concept of any other user. We consider all such matching users for our work. Once such users are found out, we calculate the user who has the highest number of matching concepts with the current user is chosen and the interested pages for the current query for the finally chosen user will be returned. In case more than one user is having the highest number of matching concepts, we can resolve this by selecting either of them in a random fashion. This method is good for a majority of queries but doesn't work for common interesting concepts under widely covered topics. Figure. 4. shows us how dynamic groups are formed. The Current user is not having the queries query1 and query 2 as interested query concepts in the user-interest tree. For the same user, for Query1 the user may be grouped with users 4, 5, and 6. Users 4, 5 and 6 are considered for this query because they have query 1 as an interested concept. The interesting concepts of User5 is given as output for this query for the current user because this is the user who has the most number of matching concepts with the current user and for Query2, it is User2 whose interesting concepts are considered interesting concepts of current user since of all the users who have the Query2 as an interested concept, it is User2 who has the maximum number of matching concepts with the current user.

4 Results and Evaluation

We implemented the above variations of the Hits algorithm and tested the performance of our methodologies with a set of one million Tamil documents for the tourism domain. We obtained these documents by crawling the net with tourism specific seed URLs. We have also compared our modified HITS algorithms with the existing UNL based ranking methodology as the baseline for the same set of documents, with 93 tourism specific queries. The evaluation was done by a set of 20 students working in the area of Information Retrieval in the University's Computer Science and Engineering Department. We

consider CoRee as our baseline for evaluation and not the traditional HITS algorithm since HITS was designed to specifically suit keyword based search whereas CoRee is a concept based search engine. Though CoRee, is a search engine for the Tamil language, our methodologies are not language specific. They can very well be used for any other language. The underlying language independent UNL representation of the documents and the concepts in the user profiles makes this task an easier one. The evaluation parameters used here are precision@5 and precision@10.

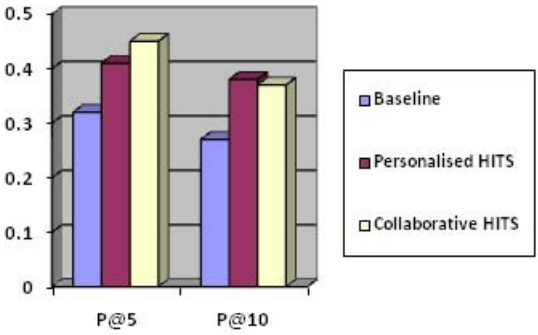


Fig. 5. Evaluation - P@5 and P@10

Figure. 5 shows the performance of both our new methodologies against the baseline. The results indicate that there is a remarkable improvement in the P@5 and the P@10 values by our new methodologies than the baseline. This improvement is attributed to combining semantics with personalization and collaboration. We can also see from the figure that personalization is better at times and collaboration is better at different times. We found that when there is a clear drift in the user interests, it is been reflected in the search results. Another important result that we came across is the effect of cold start users in the case of personalized search. Cold start users are those who come into the group as first time users which means that they have no user profiles. The first session of search could prove to be not very fruitful with the personalized HITS and has a high chance of resembling the baseline results. These cold start users affect the performance of our system considerably. This problem is handled in collaboration since a user's search depends on other user profiles and not solely on his own user profile. Apart from this problem, we found that the user profile creation is dynamic and hence time consuming. Though there are methodologies that statically create user profiles and hence are less time consuming, we opted for dynamic grouping so as to handle the user interest matches dynamically so that the most recent changes in user interests are being captured. However, we would like to create profiles statically and do the ranking as an offline process as well and study the trade-off between the efficiency and effectiveness of the system as our future work. Apart from offline profile creation, we are currently in the process of developing an offline ranking algorithm that has a reduced time consumption. The algorithm that we chose to modify for this purpose is the popular PageRank algorithm.

4.1 Conclusion

In this paper, we have proposed a new variation to the popular HITS algorithm so as to adapt it to Personalizing the web search process. We have also proposed a new data structure for efficient storage and retrieval of the User interested concepts. We have explained two variations to the collaborative search process, one in which all the users are considered to belong to the same interest group and the other variation, which gives the innovative idea of the dynamic group formation considering the fact that the current user has a similar interest to some user for the current query only and may not be interested as the same person for a different query. We would like to improve the retrieval time of the collaborative search process by forming offline groups based on the similarities in the concepts users are interested in as our future work. We are currently in the process of developing modified versions of the offline PageRank algorithm with the intention of reducing online time consumption.

References

1. Delgado, J., Ishii, N., Ura, T.: Content-based collaborative information filtering: Actively learning to classify and recommend documents. In: Klusch, M., Weiss, G. (eds.) CIA 1998. LNCS (LNAI), vol. 1435, pp. 206–215. Springer, Heidelberg (1998)
2. Hu, J., Chan, P.K.: Personalized Web Search by Using Learned User Profiles in Re-ranking. In: Proceedings of Web KDD, Workshop on Web Mining and Web Usage Analysis (2008)
3. Pérez, J.D.J., Calderón, M.L., González, C.N.: Towards an Information Filtering System in the Web Integrating Collaborative and Content Based Techniques. In: IEEE Proceedings of the First Latin American Web (2003)
4. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (1997)
5. Mittal, N., Nayak, R., Govil, M.C., Jain, K.C.: A Hybrid Approach of Personalized Web Information Retrieval. In: IEEE International Conference on Web Intelligence and Intelligent Agent Technology (2010)
6. Palleti, P., Karnick, H., Mitra, P.: PersonalizedWeb Search using Probabilistic Query Expansion. In: International Conferences on Web Intelligenceand Intelligent Agent Technology - Workshops (2007)
7. Roma, Y., Shtykh, Q.J.: Integrating Search and Sharing: User-Centric Collaborative Information Seeking. In: Eigth IEEE/ACIS International Conference on Computer and Information Science (2009)
8. Sendhilkumar, S., Geetha, T.V.: User Representation in Personalized Web Search using Interest Vectors. In: International Journal of Recent Trends in Engineering (2009)
9. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User Profiles for Personalized Information Access. Artificial Intelligence Publisher: Springer (2007)
10. Umamaheswari, E., Geetha, T.V., Parthasarathi, R., Karky, M.: A Multilevel UNL Concept based Searching and Ranking. In: WEBIST (2011)
11. UNL- `http://www.undl.org`
12. Shu-hong, Y., Fu-liang, W.: Study on Personalized Search Engine Based on Files. In: IEEE International Conference on Internet Technology and Applications (2010)
13. Zeng, J., H.-J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behaviour. In: Proceedings of the 16th International Conference on World Wide Web, pp. 151–160 (2007)

14. Shi, X.: An intelligent knowledge-based recommendation system. Intelligent information processing II, 431-435 (2005)
15. Jeon, H., Kim, T., Choi, J.: Adaptive User Profiling for Personalized Information Retrieval. In: Third International Conference on Convergence and Hybrid Information Technology (2008)
16. Miller, J.C., Rae, G., Schaefer, F., Ward, L.A., Lofaro, T., Faraha, A.: Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 444–445 (2001)
17. Jayanthi, J., Jayakumar, K.S.: An Integrated Page Ranking Algorithm for Personalized Web Search. International Journal of Computer Applications 12 (2011)
18. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intelligence and Agent Systems: An International Journal, 219–234 (2003)
19. Sendhilkumar, S., Geetha, T.V.: Concept based Personalized Web Search. Advances in Semantic Computing 2, 79–102 (2010)
20. Daoud, M., Tamine-Lechani, L., Boughanem, M.: Using A Concept-based User Context For Search Personalization. In: Proceedings of the World Congress on Engineering, vol. I (2008)
21. Leung, K.W.-T., Lee, D.L., Ng, W., Fung, H.Y.: A Framework for Personalizing Web Search with Concept-Based User Profiles. ACM Transactions on Internet Technology 2(3) (2001)
22. Prabaharan, S., Wahidabanu, R.S.D.: Ontological Approach for Effective Generation of Concept Based User Profiles to Personalize Search Results. Journal of Computer Science, 205–215 (2012)

Group Recommender System Based on Rank Aggregation – An Evolutionary Approach

Ritu Meena and Kamal K. Bharadwaj

School of Computer and System Science
Jawaharlal Nehru University Delhi, India
{meena.ritu,kbharadwaj}@gmail.com

Abstract. Recommender systems (RSs) have emerged as a remarkable tool that very effectively handle information overload problem caused by unprecedented growth of resources available on the www. RSs research has mainly focused on algorithms for recommending items for individual users. However, Group recommender systems (GRSs) provide recommendations to group of persons i.e. they take all individual group members' preferences into account and try to satisfy them optimally. The well known Kemeny optimal aggregation generates an aggregated list that minimizes the average Kendall tau Distance from the input lists; however such aggregation is NP-Hard. In this work, we design and develop a novel approach to GRS based on Kemeny optimal aggregation using genetic algorithm (GA). We have employed edge recombination operator (ERO) and scramble sub-list mutation as genetic sequencing operators. Experimental results clearly demonstrate that proposed GA approach to rank aggregation (RA) based GRS, GA-RA-GRS outperforms the well known GRS techniques.

Keywords: Group Recommender Systems, Genetic Sequencing Operator, Edge Recombination Operator, Mutation, Kendall Tau Distance.

1 Introduction

Web based Recommender Systems (RSs) [1] are the most illustrious application of the web personalization to deal with problems of information and product overload. RSs help online consumers by providing suggestions that effectively prune large information spaces so that users are directed toward those items that best meet their needs and preferences. The interest in such systems has dramatically increased due to the demand for personalization technologies by large and successful e-Commerce platforms. With the explosive growth of resources available through the internet, information overload has become a serious concern. Since their invention they have been used for recommending books, CDs, movies, jokes, news, electronics, travels and many other products and services- some well known RSs include Amazon.com., MovieLens, Netflix, Jester etc.

Most of the research in the area of RSs has focused on algorithms for recommending items for individual users. However, in certain domains it may be desirable to be able to recommend items for a group of persons e.g. movies, restaurants etc. PolyLens was the first known group recommender system (GRS) for

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 663–676, 2013.
© Springer International Publishing Switzerland 2013

recommending movies [11]. Thereafter several successful GRSs have been developed- some well known GRS include MUSICX (music), INTRIGE (tourism), YU'S TV RECOMMENDER (television program) etc.

Group recommender systems (GRSs) [3], [14] provide recommendations to groups i.e. they take all individual group members' preferences into account and satisfy them optimally with a sequence of items or a single item recommendation. In general optimal solution is not possible for GRSs and several techniques proposed in the past try to generate near optimal solution. No unique solution is possible for GRS in general.

Group recommender systems (GRSs) [7] can be classified into two main categories: *aggregated models*, which aggregate individual user data into a group data, and generate predictions, based on the group data; and *aggregated predictions*, which aggregate the predictions for individual users into group predictions. We have considered the way in which individual preferences are obtained (by content-based or collaborative filtering) as an additional dimension to be taken into account in such categorization. In any of the above cases, the mechanisms in which user profile models or item predictions are aggregated are manifold, and can be based on any of the social choice strategies.

In order to generate effective recommendations for a group the system must satisfy, as much as possible, the individual preferences of the group's members. Group recommendation approaches are either based on the generation of an integrated group profile or on the integration of recommendations built for each member separately.

The rank aggregation (RA) problem is to combine many different rank orderings on the same set of candidates, or alternatives, in order to obtain a "better" ordering. RA has been studied extensively in the context of social choice strategy where several "voting paradoxes" have been discovered. The problem also arises in many other settings.

In GRSs [7] different strategies to combine several users' preferences and to aggregate item ranking lists can be applied based on the utilized social welfare function. These strategies are classified by three categories:

***Majority-Based Strategies*: w**hich strength the "most popular" choices (user preferences, item rankings, etc.) among the group, e.g. Borda Count, Copeland Rule, and Plurality Voting strategies.

***Consensus-Based* (or *Democratic*) Strategies:** which average somehow all the available choices, e.g. Additive Utilitarian, Average without Misery, and Fairness strategies?

***Border-Line* Strategies:** also called *role-based* strategies in, which only consider a sub-set of choices based on user roles or any other relevant criterion, e.g. Dictatorship, Least Misery and Most Pleasure strategies.

Rank Aggregation (RA) strategy has been successfully used in the area of GRSs (e.g. Least misery, Most pleasure, Borda count and Copeland rule. [7].

Among various aggregation strategies, Kemeny optimal aggregation [4] is considered as unique method that stimulatingly satisfy neutrality natural, important and consistency properties referred to in the social choice literature. Kemeny optimal aggregation produces "best" compromise ordering that minimizes the average Kendall tau distance (KtD) from the input orderings. However, such aggregation is NP-hard [15].

Genetic algorithms (GAs) [10] are search algorithms that are determined by the mechanics of natural selection and natural genetic. Holland (1975) introduced a method of studying natural selection and Mendelian genetics. In the GA approach, each point in the search space (or population), is a set of chromosomes which represents a possible solution. This approach requires a population of chromosome, which representing a combination of features from the set of features and require a cost function that calculates each chromosome's fitness (this function is called evaluation function or fitness function). The algorithm performs optimization by manipulating a finite population of chromosomes. In each generation the GA creates a set of new chromosomes by crossover, inversion and mutation, which correlate to processes in natural reproduction.

Genetic algorithms (GAs) are well established techniques to produce near optimal solution to difficulties and time consuming NP-hard problems.

In this paper, we have designed and developed a genetic algorithm approach to Kemeny optimal RA based GRS.

The rest of the paper is organized as follows: Section 2 provides an overview into the group recommender system (GRS) survey, rank aggregation (RA) problem, genetic algorithms (GA) and genetic sequencing operates. The proposed GA approach to RA based GRS is introduced in Section 3. Details of the experiments performed and the results so obtained are given in Section 4. Finally, Section 5 presents conclusions and points out some directions for future research.

2 Background

Recommender systems [5] are a step towards the new paradigm of "item searching for a user" rather than the other way round. The explosive growth of the web [2] has led to "information overload", the overwhelming plethora of choices and options available to a user, which may also vary widely in quality. RSs are personalization tools which enable users to be presented information suiting his interests, which are novel, serendipitous and relevant, without being explicitly asked for.

The extension of traditional recommendation algorithms to recommending items to members of a group is not straightforward and the different issues need to be identified and resolved.

2.1 Group Recommender Systems (GRSs)

Though recommendation approaches have addressed group preference modeling explicitly to a rather limited extent, or in an indirect way in prior work in the computing field, the related issue of *social choice* (also called *group decision making*, i.e. deciding what is best for a group given the opinions of individuals) has been studied extensively in Economics, Politics, Sociology, and Mathematics. The models for the construction of a social welfare function in these works are similar to the group modeling problem we put forward here [7].

Approaches to GRSs are classified into the ones that make the group Recommendation out of the individuals' recommendations, and those that merge the

profiles of multiple people and create a single group preference model. In the past, rank aggregation has been very successfully utilized to develop various strategies for GRS [7], [8]. Some well known strategies are discussed below (Fig. 1– Fig. 4).

- **Least Misery Strategy.** It takes the minimum of individual ratings. Item B's group rating is 4, namely the smallest of 4, 9, and 5. Example. *PolyLens*.

Items	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
$\mathbf{u_1}$	8	10	4	3	6	10	9	6	8	10
$\mathbf{u_2}$	8	1	9	8	9	7	9	6	9	3
$\mathbf{u_3}$	6	10	5	2	7	9	8	5	6	7
group	**6**	**1**	**4**	**2**	**6**	**7**	**8**	**5**	**6**	**3**

Fig. 1. Least misery strategy. Rank list of items for the group ($i_{7,}$ $i_{6,}$, i_1-i_5-$i_{9,}$, $i_{8,}$ $i_{3,}$ $i_{10,}$ $i_{4,}$ i_2).

- **Most Pleasure Strategy.** It takes the maximum of individual ratings. B's group rating is 9, namely the largest of 4, 9 and 5.

items	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
$\mathbf{u_1}$	8	10	4	3	6	10	9	6	8	10
$\mathbf{u_2}$	8	1	9	8	9	7	9	6	9	3
$\mathbf{u_3}$	6	10	5	2	7	9	8	5	6	7
group	**8**	**10**	**9**	**8**	**9**	**10**	**9**	**6**	**9**	**10**

Fig. 2. Most pleasure strategy. Rank list of items for the group (i_2- i_6- i_{10}, i_3-i_5-i_7- $i_{9,}$ i_1-$i_{4,}$ i_8).

- **Borda Count Strategy.** It counts points from items' rankings in the individuals' preference lists, with bottom item getting 0 points, next one up getting one point, etc. A's group rating is 17, namely 0 (last for user2) + 9 (first for user3) + 8 (shared top 3 for user1).

U	$\mathbf{i_1}$	$\mathbf{i_2}$	$\mathbf{i_3}$	$\mathbf{i_4}$	$\mathbf{i_5}$	$\mathbf{i_6}$	$\mathbf{i_7}$	$\mathbf{i_8}$	$\mathbf{i_9}$	$\mathbf{i_1}$
u	8	1	4	3	6	1	9	6	8	1
u	8	1	9	8	9	7	9	6	9	3
u	6	1	5	2	7	9	8	5	6	7

⇩

User	$\mathbf{i_1}$	$\mathbf{i_2}$	$\mathbf{i_3}$	$\mathbf{i_4}$	$\mathbf{i_5}$	$\mathbf{i_6}$	$\mathbf{i_7}$	$\mathbf{i_8}$	$\mathbf{i_9}$	$\mathbf{i_{10}}$
$\mathbf{u_1}$	4.5	8	1	0	2.5	8	6	2.5	4.5	8
$\mathbf{u_2}$	4.5	0	7.5	4.5	7.5	3	7.5	2	7.5	1
$\mathbf{u_3}$	3.5	9	1.5	0	5.5	8	7	1.5	3.5	5.5
group	**12.5**	**17**	**10**	**4.5**	**15.5**	**19**	**20.5**	**6**	**15.5**	**14.5**

Fig. 3. Borda count strategy. Rank list of items (i_7, i_6, i_2, i_5- $i_{9,}$ i_{10}, i_1, i_3, i_8, i_4).

- **Copeland rule strategy.** It counts how often an item beats other items (using majority vote) minus how often it loses. F's group rating is 5, as F beats 7 items (B,C,D,G,H,I,J) and looses from 2 (A,E).
  Copland Index = no. of pair wise victories – no. of pair wise defeats

User	i_{10}	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_1	8	10	4	3	6	10	9	6	8	10
u_2	8	1	9	8	9	7	9	6	9	3
u_3	6	10	5	2	7	9	8	5	6	7

⇩

User	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
i_1	-	0	-	-	-	0	-	-	-	0
i_2	+	+	0	-	+	+	+	0	+	+
i_3	+	+	+	0	+	+	+	+	+	+
i_4	-	+	-	-	0	+	+	-	0	0
i_5	-	0	-	-	-	0	-	-	-	-
i_6	-	+	-	-	-	+	0	-	-	-
i_7	+	+	0	-	+	+	+	0	+	+
i_8	-	+	-	-	0	+	+	-	0	0
i_9	-	0	-	-	0	+	+	-	-	-
i_{10}	0	+	-	-	+	+	+	-	+	+
group	**-3**	**+7**	**-6**	**-9**	**+1**	**+8**	**+5**	**-6**	**0**	**+3**

Fig. 4. Copeland rule strategy. Rank list of items (i_6, i_2, i_7, i_{10}, i_5, i_9, i_1, i_3, i_8, i_2).

2.2 Rank Aggregation

A natural step toward aggregation was taken by Kemeny [4]. Informally, given k orderings r_1,……,r_k on (partial list of) alternatives {1, 2, …., n}, a *Kemeny optimal* ordering σ minimizes the sum of the Kendall tau Distance(Sum-KtD).

$$\sum_{i=1}^{k} K(\sigma, ri) \tag{1}$$

Although, Kemeny optimal solutions are likely to generate "best" compromise orderings, finding a Kemeny optimal aggregation is NP- hard. The optimal solution in general is not possible. We can only find near optimal solution.

2.3 Genetic Algorithm

A genetic algorithm approach processes a population of chromosomes where each chromosome represents a potential solution to an optimization problem. An objective

function evaluates the quality of a solution, which is called a fitness function. Usually individuals with high fitness are selected as parents and genetic operators such as crossover and mutation are applied to produce new offspring. The good newly generated individuals replace the current bad individuals to form the new population for the next generation. Suitable stopping criteria, such as, if a maximum number of generations elapses or a desired level of fitness is reached is applied to terminate GA.

Each chromosome is made up of a sequence of genes from a certain alphabet. A binary vector representation chromosome may not work well for optimization problems involving several parameters for the most natural representation of solutions e.g. sequence of nodes, real valued encoding tree etc.

GA approach has been shown to be very successful strategy to produce near optimal solutions for NP-Hard problem e. g. TSP [16]. The main steps of GA are given below:

Step1. assuming an appropriate encoding for a chromosome such that it represents possible solution to the specific problem.

Step2. Each individual's fitness is evaluated using a function that estimates its ability to solve the specified problem at hand.

Step3. Selection: Individual population members applying genetic operators (Crossover and mutation) to be parents.

Step4. Recombination: produce offspring by applying genetic operators (crossover and mutation add them to the population.

Step5. Apply fitness function to the offspring's produced at step 4 to evaluate their fitness.

Step6. Repeat steps 3-5 until the stopping criteria satisfied.

3 Proposed Group Recommender System

In this chapter, a genetic algorithm (GA) approach to rank aggregation (RA) based group recommender system (GA-RA-GRs) is presented.

3.1 Group Recommender System and Encoding

Let us consider a group of n users and m items i.e.

Group = $\{\mathbf{u_1}, \mathbf{u_2}, \ldots\ldots, \mathbf{u_n}\}$

Items = $\{\mathbf{i_1}, \mathbf{i_2}, \ldots\ldots\ldots, \mathbf{i_m}\}$

Assuming that all the users in the group have rated all the items, the rating matrix user × item is generated such that the i^{th} row represents the rating of $\mathbf{u_i}$ for $\mathbf{i_1}$, to $\mathbf{i_m}$ **e.g.** for n = 2 and m = 10 the user × item matrix would be :

it	$\mathbf{i_1}$	$\mathbf{i_2}$	$\mathbf{i_3}$	$\mathbf{i_4}$	$\mathbf{i_5}$	$\mathbf{i_6}$	$\mathbf{i_7}$	$\mathbf{i_8}$	$\mathbf{i_9}$	$\mathbf{i_{10}}$
$\mathbf{u_1}$	2	4	7	1	6	8	1	9	6	8
$\mathbf{u_2}$	5	7	3	4	9	1	9	4	6	8

We then construct a new matrix such that each row represents a set of items arranged in the decreasing order of preference of the user. Each row is a permutation of m items.

Restructuring the Permutation

Rows represent item preferences of different users in decreasing order and chromosomes would be represented as a permutation of items.

From the above user × item matrix we generate the following data set such that i^{th} row represents the sequence of items in the decreasing order of preference for the user $\mathbf{u_i}$, list the permutation of items.

The GRS problem is now to generate a permutation of 10 items in aggregating the n permutation that satisfies n users optimally.

Encoding

A chromosome is simply a permutation of items i. e. (i_7, i_8,…. i_4) which is encoded as (7, 8, ….4). For example,

it	$\mathbf{i_1}$	$\mathbf{i_2}$	$\mathbf{i_3}$	$\mathbf{i_4}$	$\mathbf{i_5}$	$\mathbf{i_6}$	$\mathbf{i_7}$	$\mathbf{i_8}$	$\mathbf{i_9}$	$\mathbf{i_{10}}$
$\mathbf{u_1}$	i_7	i_8	i_6	i_{10}	i_3	i_9	i_5	i_2	i_1	i_4
$\mathbf{u_2}$	i_7	i_6	i_1	i_2	i_9	i_1	i_4	i_8	i_3	i_5

is represented as:

Ite	$\mathbf{i_1}$	**i**	$\mathbf{i_3}$	$\mathbf{i_4}$	$\mathbf{i_5}$	$\mathbf{i_6}$	$\mathbf{i_7}$	**i**	$\mathbf{i_9}$	$\mathbf{i_{10}}$
$\mathbf{u_1}$	7	8	6	1	3	9	5	2	1	4
$\mathbf{u_2}$	7	6	1	2	9	1	4	8	3	5

GRS finally recommends the best chromosome (minimum Sum- Ktd) i.e. the permutation with the minimum Sum-KtD.

3.2 Fitness Function(Sum-KtD)

Consider a list of items L = {1,…, n} and suppose that σ_1 and σ_2 are partial lists of L. Then K (σ_1, σ_2) is defined as:

K (σ_1, σ_2) = the number of pairs (i, j) in (1….n) such that $\sigma_1(i) < \sigma_1(j)$ *but* $\sigma_2(i) > \sigma_2(j)$

This is to be noted that in case

(i) Both i and j do not appear in lists $\sigma 1$ and $\sigma 2$, then the pair (i, j) contributes nothing to the K (σ_1, σ_2).

(ii) For any two partial lists if K (σ_1, σ_2) = K (σ_2, σ_1) for σ_1 and σ_2 are partial.

(iii) If we assume that σ_1 and σ_2 are full lists (permutations) then K (σ_1, σ_2) is a metric known as the Kendall tau distance between them and

In our GRS system we are assuming σ_1 *and* σ_2 as full list.

Here is a simple example of this problem. Suppose we receive the resulting propriety lists (of size 10) from 5 different search users and wish to form a consensus list using the distance method described above.

Sum-KtD. For a collection of partial lists $r_1, r_2, \ldots, r_n$ and a full list σ we define sum of Kendall tau distance (Sum-KtD) as:

$$Sum - KtD(\sigma, r_1, r_2, \ldots, r_n) = \sum_{i=0}^{n} K(\sigma, r_i) \quad (2)$$

Given a set of preference lists or rankings, we wish to find a consensus that minimizes the Kendall Tau distance (i.e., find a Kemeny optimal aggregation). The Kendall Tau Distance (i.e., n permutations) is the number of pairs of distinct integers σ_i and σ_j such that $1 \leq i, j \leq n$ where σ_i and σ_j are in opposite order in each ranking. for example, if n = 10 and σ is 4 1 2 9 7 8 3 5 6 10 while R1 is 5 1 3 2 4 6 7 9 10 8, we find the Kendall Tau distance to be 19 since only the 19 pairs {4, 5}, {4, 1}, {4, 3}, {4, 2}, {1, 5}, {2, 5}, {2, 3}, {9, 7}, {9, 3}, {9, 5}, {9, 6}, {7, 3}, {7, 5}, {7, 6}, {8, 3}, {8, 5}, {8, 6}, {8, 10} and {3, 5} are in opposite order in each ranking. This is a standard method used by mathematicians to quantify the difference (i.e., distance) between the two rankings. An example to explain computation of fitness function as Sum-KtD is given in Table 1.

Table 1. Fitness Function (Sum-KtD).

σ = 4 1 2 9 7 8 3 5 6 10

Ranking		Ktd
R1	5 1 3 2 4 6 7 9 10 8	19
R2	3 10 4 8 1 7 2 6 5 9	23
R3	4 8 5 10 9 7 6 3 1 2	24
R4	5 4 10 6 7 8 9 3 2 1	31
R5	4 6 7 10 8 5 3 2 9 1	28
	Total distance(Sum-KtD)	**125**

3.3 Selection Criteria

We are using elitism**,** where best individuals are retained from generation to generation and this would avoid possibility of crossover or mutation destroying such individuals.

3.4 Offspring Generation

In our GA based scheme, we have employed the following genetic sequencing operator to generate offspring:

Edge Recombination Operator (ERO)

The essence of edge recombination is to achieve maximal inheritance from parental edge. Its ability to preserve parental edges has been validated through its higher correlation coefficient between the fitness of parents and offspring [16]. The main steps of ERO are as under:

Step1. Choose the initial item from one of the two edges. (It can be chosen randomly or according to criteria outlined in step 4.) This is the "current item".

Step2. Remove all occurrences of the "current item" from the left- hand side of the edge map.(these can be found by referring to the edge list for the current item.)

Step3. If the current item has entries in its edge list go to step 4; otherwise, go to step 5.

Step4. Determine which of the items in the edge list of the current item has the fewest entries in the own edge-list. The item with the fewest entries becomes the "current item." Ties are broken randomly. go to step 2.

Step5. If there is no remaining "unvisited" item, then STOP. Otherwise, randomly choose an "unvisited" item and go to step 2.

An example of ERO is given in Fig. 5.

Example : **Parent1: 2 3 4 5 7 1 8 9 10 6**

Parent2: 4 1 3 5 7 8 10 6 2 9

Edgelist:

2	3 , 6, 9
3	2, 4, 1, 5
4	3, 5, 9, 1
5	4, 7, 3
7	5, 1, 8
1	7, 8, 4, 3
8	1, 9, 7, 10
9	8, 10, 2, 4
10	9, 6, 8
6	10, 2

OFFSPRING: 2 6 10 9 8 7 5 4 3 1

Fig. 5. Edge Recombination operator (ERO).

Scramble Sub-list Mutation

A sub-list of items is randomly selected in list of ordered items representing the parent chromosome and scrambled leaving the rest of the chromosome unchanged [9] as illustrated in Fig. 6.

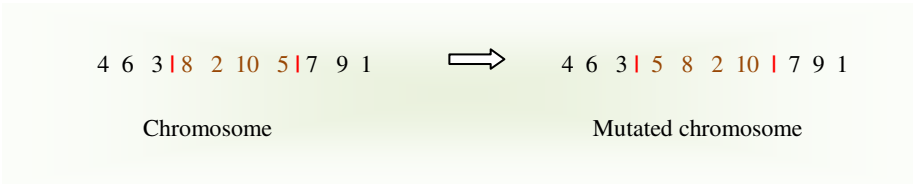


Fig. 6. Example of scramble sub list mutation

3.5 Stopping Criteria for Genetic Algorithm

Evolution process is terminated when there is no improvement in the best solution for 20 consecutive generations.

The main steps of our proposed GA-RA-GRS are summarized as under:

Step1. Initial Population

Set of randomly generated permutations.
Evaluate the fitness value of the individuals using Sum-KtD

Step2. Offspring Generation

(i) Crossover
Two parent chromosomes are selected and new offspring is generated using ERO.

(ii) Mutation
A parent chromosome is selected and offspring is generated as mutated chromosome using scramble sublist mutations.

Evaluate the fitness value of the individuals using Sum-KtD

Step3. Stopping criteria

If *stopping condition* = true
then return the best individual as the solution and STOP
Else go to Step 2

Step4. Finally our proposed GRS recommends the permutation with the least Sum-KtD

The block diagram of the proposed GA-RA-GRS is presented in Fig. 7.

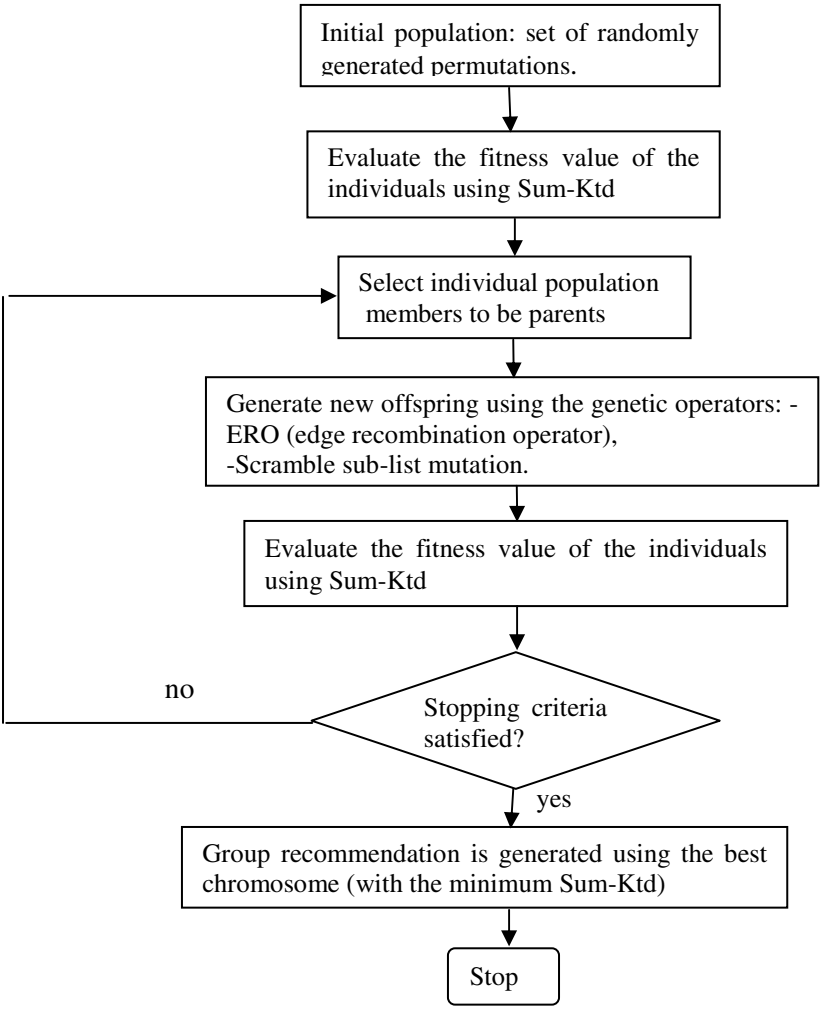


Fig. 7. Block diagram of the proposed GA-RA-GRS.

4 Experiments and Results

We have conducted experiments to evaluate performance of our proposed GA-RA-GRS as compared to different baseline GRS techniques on the basis of sum of Kendall Tau distance (Sum–KtD) - smaller the Sum-KtD better the technique. For our experiments, we have considered fixed size recommendation list of 20 items (chromosomes as permutations of 20 items) with different group sizes, Group5, Group10 and Group15. GA parameters are chosen as: population size 20; crossover probability: 0.8; mutation probability: 0.2.

4.1 Experiment 1

In this experiment, Sum- KtD is computed across generations for three different group sizes as shown in Table 2 and depicted in Fig. 8 that clearly shows that the GA converges to near optimal solution after 1600 generations for Group5, Group10 and Group15.

Table 2. Sum of Kendall Tau Distance (Sum-KtD) across generations for different group sizes

Generation	Group5	Group10	Group15
0	381	813	1237
200	365	699	1092
400	359	655	1041
600	354	653	970
800	353	647	960
1000	350	639	956
1200	349	639	955
1400	347	633	955
1600	346	631	955
1800	346	631	955
2000	346	631	955

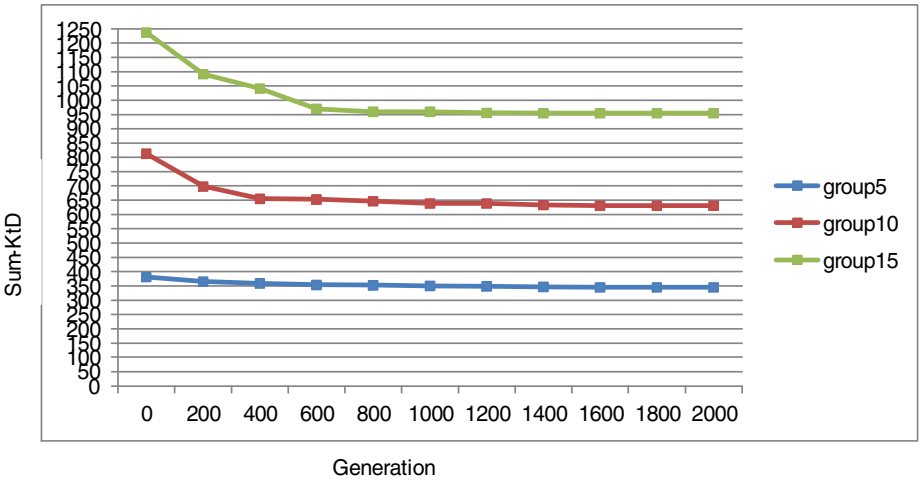


Fig. 8. Variation of Sum-Ktd across generations for different group sizes

4.2 Experiment 2

Here, we have compared the proposed GA-RA-GRS with the well known GRS techniques. The results of the experiments are summarized in Table 3 and depicted in Fig. 9. That clearly demonstrates that the proposed GR-RA-GRS outperforms Least misery, Most pleasure, Borda count and Copeland rule.

Table 3. Comparison of GA-RA-GRS with well known GRS techniques

Group size	Least misery	Most pleaser	Borda-count	Copeland rule	GA-RA-GRS
5	467	480	450	456	346
10	867	893	891	840	631
15	1268	1402	1439	1412	955

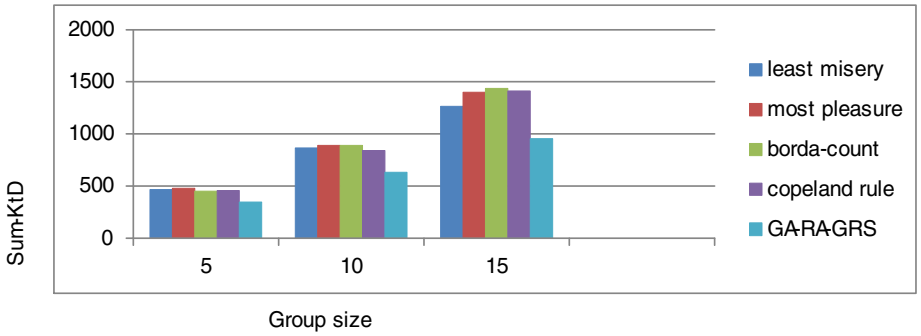


Fig. 9. Effectiveness of proposed GA-RA-GRS as compared to baseline techniques for different group sizes

5 Conclusions and Future Work

In this paper, a frame work for group recommender system (GRS) is presented based on Kemeny optimal aggregation- a NP hard problem. Genetic algorithm (GA) is successfully employed to generate near optimal solution. Experimental results are quite encouraging and demonstrate that the proposed GA-RA-GRS outperforms several baseline GRS techniques.

Our future work will focus on considering modified EROs [6], [12] to develop efficient genetic sequencing operators to further improve the proposed GRS approach. Regarding Kemeny optimal aggregation, we have assumed full sub-lists permutations in our work. In this regard, consideration of partial sub-list in the framework of GRS needs further investigation. The computational complexity is a major problem for the proposed GA-RA-GRS and therefore a parallel genetic algorithm approach would be an interesting future work. It is to be seen how the ideas presented in this work can be utilized for effective negotiation mechanism [13].

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734–749 (2005)

2. Anand, D., Bharadwaj, K.K.: Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities. Expert Systems with Applications. Elsevier (2010) (in press)
3. Baltrunas, L., Makcinskas, T., Ricci, F.: Group Recommendations with Rank Aggregation and Collaborative Filtering. In: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010), pp. 119–126 (2010)
4. Baskin Jacob, P., Krishnamurthi, S.: Preference aggregation in group recommender systems for committee decision-making. In: RecSys 2009, pp. 337–340 (2009)
5. Bharadwaj, K.K., Al-Shamri, M.Y.H.: Fuzzy-Genetic Approach to Recommender Systems Based on a Novel Hybrid User Model. Expert Systems with Applications 35, 1386–1399 (2007)
6. Ting, C.-K.: Improving Edge Recombination through Alternate Inheritance and Greedy Manner. Evo COP 2004, 210–219 (2004)
7. Cantador, I., Castells, P.: Group Recommender Systems: New Perspectives in the Social Web. In: Pazos Arias, J.J., Fernández Vilas, A., Díaz Redondo, R.P. (eds.) Recommender Systems for the Social Web. Intelligent Systems Reference Library, vol. 32, pp. 139–158. Springer, Heidelberg (2012)
8. Garcia, I., Pajares, S., Sebastia, L., Onaindia, E.: Preference elicitation techniques for group recommender systems. Information Sciences 189, 155–175 (2012)
9. Lawrence, D.: Schedule Optimization Using Genetic Algorithms. In: Handbook of Genetic Algorithms, Van Nostr and Reinhold, New York (1991)
10. Mitchell, M.: An introduction to genetic algorithms, pp. I-VIII, 1–208. MIT Press (1998) ISBN 978-0-262-63185-3
11. O'Connor, M., Cosley, D., Konstan, J., Riedl, J.: PolyLens: A recommender system for groups of users. In: Proceedings of the European Conference on Computer-Supported Cooperative Work (2001)
12. Nguyen, H.D., Yoshihara, I., Yasunaga, M.: Modified Edge Recombination Operators of Genetic Algorithm for the Travelling Salesman Problem. In: Proc. IEEE Int. Conf. on Industrial Electronics, Control, and Instrumentation (2000)
13. Eva, O., García, I., Sebastia, L.: A Negotiation Approach for Group Recommendation. In: Proceedings of the International Conference on Artificial Intelligence (ICAI-2009), 919–925. CSREA Press (2009)
14. Herr, S., Rösch, A., Beckmann, C., Gross, T.: Informing the design of group recommender systems. CHI Extended Abstracts (2012)
15. Sivakumar, D., Dwork, C., Kumar, R., Naor, M.: Rank aggregation methods for the Web, WWW 2001 (2001)
16. Whitley, D., Starkweather, T., Fuquay, D.: Scheduling problems and travelling salesman: The genetic edge recombination operator. In: International Conference on Genetic Algorithms, pp. 133–140 (1989)

Concept Similarity Based Academic Tweet Community Detection Using Label Propagation

G. Manju and T.V. Geetha

Anna University, Chennai, TamiNadu, India
manju.shruthi@gmail.com, tvg@hotmail.com

Abstract. In today's world, Social Network plays a vital role in the society. Social Network users share their ideas, views, opinions, and develop their personal relationship. Social Network has major influence with academic community. This paper aims at detecting similar concept based academic tweets from the numerous available tweets and forming a community, considering the social relation between the tweeters. Academic community can support recommender system for researcher network. In our work, in order to extract concept similarity based academic community, concept similarity graph is constructed from twitter. Label Propagation algorithm is used to detect academic community. Normally, tweets contain user views, suggestions and discussion on a specific topic. In spite of tweets, containing other words in it, Concept words play a vital role in identifying about the aim of the tweeter in posting the tweet. Moreover, for academic topics, academic concepts are important. So, the Concepts are extracted and based on the similarity between concepts, academic community has been extracted from twitter. Label propagation has proven to be an effective method for detecting communities in complex networks. In this work, the new update rule based on social relation is introduced for Label propagation algorithm and used for concept based community detection. The experiment shows that, in comparison with standard label propagation algorithm, the label propagation with modified update rule reduces the number of iterations for convergence and as well was more effective in detecting communities.

Keywords: Social Network Analysis, Label Propagation, Community detection, Semi-supervised learning, String Kernel, Tweet, Concept Similarity.

1 Introduction

Twitter is the leading micro blogging social network. It is the ninth most popular site on the Internet with over 200 million registered users producing over 200 million tweets every day. Users post publicly viewable tweets of up to 140 characters in length, and follow other users whose tweets they are interested in receiving. Thousands of scholars and higher education institutions are participating in social media (such as Twitter), as an important aspect of their research and teaching work. Most scientific organizations, newspapers, and science journals are on Twitter, and by following them you have an up-to-date news stream about their activities. The sheer volume of data produced by Twitter makes it an attractive area of study for machine learning. Unfortunately, many

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 677–686, 2013.
© Springer International Publishing Switzerland 2013

standard algorithms for extracting information from a body of text assume correct English. As a result, they are ineffective at analyzing tweets, which often contain slang, acronyms, or incorrect spelling or grammar. Content analysis on Twitter poses unique challenges: posts are short (140 characters or less) with language unlike the standard written English on which many supervised models in machine learning and NLP are trained and evaluated. The tweet discuss on different topics. The tweets may contain academic related concepts and such tweets are named as academic tweets. Identification of academic tweets and grouping the similar academic tweets will support the researchers and academicians for their academic activities.

Our work aims in extracting academic tweets with similar concepts, and forming a community with individual tweet concepts as node and the edge representing the similarity measure between the nodes.

2 Related Work

Most of the published work in Twitter has focused on questions related to Twitter's network and community structure. For example, general features of the Twitter social network such as topological and geographical properties, patterns of growth, and user behaviours are summarized by [1].[2] argue from a network perspective that user activities on Twitter can be thought of as information seeking, information sharing, or as a social activity.

The systematic analysis of the textual content of posts on Twitter has been rarely addressed. The work carried out by [3] has examined Twitter content with respect to specific Twitter conventions: @user mentions and [4] uses Tweet hashtags to cluster into meaningful topic groups. [5] has characterized the content on Twitter and other "Social Awareness Streams" via a manual coding of tweets into categories of varying specificity, from "Information Sharing" to "Self Promotion". The other forms of content analysis on Twitter includes modelling conversations [6].The method for finding multi-level content similarity approach, inorder to detect and track the breaking news from Twitter was proposed by [7]. Since contents in microblogs have short lengths, they have emphasized on specific terms called named entities. In the first level, similarity is defined by TF-IDF. Message groups are obtained in the first level. In the second level, the author construct a network from the message groups and named entities and perform a community detection. [8] focus on automatically clustering and classifying "tweets", into predefined topic categories like, News, Sports, Entertainment, Science, Technology, Money, and "Just for Fun",based on hashtags using k-means clustering algorithm. Generally, in social networks, blog or post on some topics by some actor influence the other actors .[9] have analysed Twitter to determine the features that can help to determine the influential (or authoritive) users on a certain topic .They have considered both the history of posted messages as well as a user's position in the social graph. The goal is to better understand the concept of influence and to derive which characteristic features of users play a role when determining influence. Although rich with insight, these works do not present automatic methods for detecting academic concept community from the content of Twitter posts, the problem we approach here.

A great interest is found in identifying communities in networks. Specifically, a community in a network is a group of nodes that are similar to each other and are dissimilar

from the rest of the network. Several algorithms have been proposed which are typically classified as: divisive, agglomerative (depending on whether they focus on the addition or removal of edges to or from the network) and optimization (continuously updating the network partition in order to maximize a given measure of the quality of the network partition (i.e., the modularity). Recently [10] proposed a label propagation algorithm (LPA) for detecting network communities .This algorithm uses only the network structure as a guide,and can be summarized as follows: Each node in the network is first given a unique label; at each iteration, each node is updated by choosing the label which is the most frequent among its neighbors - if multiple choices are possible (as for example in the beginning), one label is picked randomly. [11] introduced a new form of label propagation, where a node's label propagates to neighboring nodes, according to their proximity. The labeled node acts as sources, which push out the labels through unlabeled data.[12] proposed a different label propagation algorithms that convey two unique strategies of community formation, namely, defensive preservation (preference to core of community) and offensive expansion (preference to border of community) communities is proposed-to reduce the formation of one major community. Denser networks (with higher average degrees) prefer the defensive preservation, whereas sparser networks (with lower average degrees) favor the offensive expansion of communities.[13] proposed a new update rule to reduce the computational cost and as well to improve the quality of detected communities using clustering coefficient parameter 'c'.

The work presented so far, do not address the improvement in computational cost of label update based on social relation. In this paper, we address this issue by introducing a variation of the label update rule of LPA using the social relation scoring function.

The remainder of the paper is organized as follows. Section 3 in detail describes the methodology used for tweet extraction and pre-processing, concept extraction, concept graph construction and label propagation algorithm Section 4 presents the evaluation and the section 5 deals with the concluding remarks and future enhancements.

3 Methodology

The work aims at forming a community based on similar academic concepts. The AlchemyAPI [14] was used to extract concepts from each tweet. This paper introduces an enhancement in Label propagation algorithm to detect academic community from tweets available in Twitter social network. This section describes in detail about the academic concept community detection. Fig.1 depicts the steps involved in academic tweet concept community detection.

3.1 Tweet Extraction and Pre-processing

To detect academic communities from twitter, we created a data set consisting of tweets collected through Twitter's search API [15] during the first, second and third week of April 2013. To seed our Twitter message search, we selected a total of 100 popular concepts of computer science domain that cover our 10 predefined topics, based on what was trending in the network prior to collection and queried the Twitter API for a maximum of 500 tweets per hash-tag every hour for three weeks.

As a pre-processing step, certain rules have been introduced to normalize the tweet and reduce the feature space.

- words are tokenized based on white spaces
- special characters like @ # —are removed

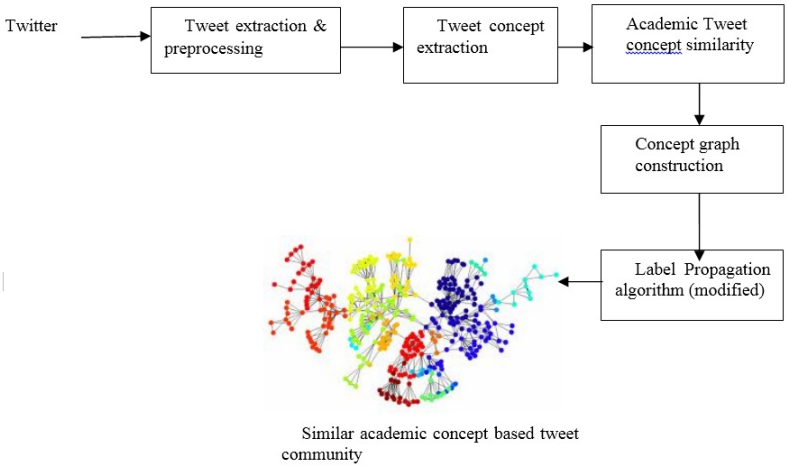


Fig. 1. Block diagram of Academic tweet concept community detection

3.2 Concept Similarity

In this work, Concept similarity is computed using String kernel function [16-17]. String Kernel is a kernel function that operates on strings to measure the similarity of pairs of strings. Using string kernels with kernelized learning algorithms such as support vector machines allow such algorithms to work with strings, without having to translate these to fixed-length, real-valued feature vectors. String kernel function has been used instead of cosine similarity(TF-IDF), because the String kernel function removes the need for stemming and lemmatization. The cosine similarity (TF-IDF) is a bag-of-words approach. In cosine similarity (TF-IDF) approach, word by word comparison is carried out to determine the similarity measure. The String subsequence kernel method computes the similarity between two strings without explicitly extracting the features. The two strings are considered more similar, the more common subsequences exist in the pair of strings. Feature space is generated by all the substrings of bounded length. A significant characteristic is substrings can be non-contiguous, and the gaps are taken into consideration. Substrings are weighted according to the degree of contiguity in a string by a decay factor λ. Moreover, tweets will be posted in languages other than English as well. If this work is enhanced to deal with tweets of languages other than English, lemmatization or stemming tools are unavailable and as well as an expensive task. As per String kernel method, string is represented as the vector and the dot product between the substrings is computed.

The string kernel method gives the similarity value between 0 to 1.We obtained the similarity value nearer to 1 if the strings are more similar and similarity value below 0.5 for non-similar strings.

3.3 Concept Graph Construction

In general Twitter, helps in learning about new research, publications, conferences, and conversations. and as well helps to interact with colleagues around the world in your own cognate fields. In specific, concepts are used in twitter feed to provide wider view about an academic topic. Hence in this work, to detect academic communities from the Twitter feeds, focus is given on concepts .We use AlchemyAPI [7] for concept extraction from Twitter feeds. AlchemyAPI provides a general implementation of Markov logic representation.

Let the network be represented as graph G(N,E),where N is the set of nodes and E is the set of edges. In this work, concept graph is constructed with node representing the concepts and the edge represents the weight between the nodes. Let the edge weight be represented as w(tc_i,tc_j),where tci and tc_j are the concept nodes and tc_i, tc_j e N. The similarity measure computed using StringKernel method is represented as edge weight.

3.4 Academic Concepts Classification

In this work, the concept based academic community is detected. The focus is given to the concepts of Computer Science domain. The extracted concepts from tweets are compared with the pre-defined list of domain concepts. The domain-specific tweet concepts are alone filtered out and used in further steps.

3.5 Label Propagation Algorithm

The community detection strategy based on label propagation introduced by [10] has proven to be an effective method for detecting communities.

Main Idea. Initially each vertex of the graph is assigned a unique label l_n where neN. During each iteration, each node adopts a new label which corresponds to the most frequent labels among its neighbors. Formally, each node n updates its label according to l_u

$$l_v = \underset{l}{argmax} \sum_{u \varepsilon N(V)} [l_u == l] \tag{1}$$

where l_v denotes the label of node 'v'

When more than one choice is possible, ties are broken randomly. This process is performed until some stop condition is met, e.g., no vertex changes its label during one step. A network community is then identified as a connected group of vertices having the same final label. The algorithm runs in linear time.

Even though LPA runs in linear time, the execution time can be improved in practice, when we process extremely large networks or move from offline to online detection. The basic idea of our improvement is to avoid unnecessary updates in each iteration,

and as well in choosing appropriate node for changing its label. As observed, at the early stage of the original LAP, most nodes are in a very diverse neighborhood. The effectiveness of updates (i.e., the fraction of attempted updates that result in changes to new labels) is high. However, the competition between communities is restricted only to their boundaries after a few iterations. For nodes inside any community, the updates are unnecessary, since they essentially do not change. As shown in [10], irrespective of size 'n', after five iterations, 95 of nodes are already correctly clustered. Additional time is required to attempt the updates that are expected to fail to change labels, so the final convergence of the algorithm is delayed. It turns out that this amount of time can be easily reduced.

Modified Label Update Rule. The propagation of a label is analogous to epidemic, idea, opinion and information spreading in a network. By assuming that a node always adopts the label of the majority of its neighbors, the LPA ignores any structures existing in this node neighborhood. This makes the algorithm very simple. However, in reality, a person adopting a new idea, often follows a neighbor who has more connections to other neighbours, because this neighbor has higher number of potential sources of information. For the same reason, when a node joining a group (i.e., changing its label to the one shared by this group) may take into account not only how many members are in this particular group (like the original LPA does) but also how well they are connected to other neighbors of the node executing the update label rule. Following this idea, we generalize the update rule of LPA as follows:

As per the standard algorithm, initially a random node is chosen as neighbour node and the label is updated based on that node. In our work, the socially related node is considered and taken as the initial neighbour node during the label update. In twitter, for a particular tweet, one or many reply tweet will exist. In our constructed concept graph, the originator tweet concept and reply tweet concepts has been related with an edge weight and they act as neighbours. During the label update, the particular tweet concept node, change its label based on the label of the reply concept nodes (socially related nodes) which act as neighbours. As per this idea, the label update rule is modified as,

$$l(i) = l(\underset{nb_u}{argmax}(score\_reply(nb_u))) \tag{2}$$

where 'l' denotes the label of a node or community i,nb_u denotes the neighbours of node i, a sub-community sharing the same label score_reply nb_u represents the scoring function of a sub-community defined as,

$$score\_reply(nb_u) = \sum_{j \epsilon nb_u \square}^{\square} (sim_{ij\square}.n(j)) \tag{3}$$

where simi,j denotes the concept similarity value between the node 'i' and node 'j' n(j) represents the number of neighbours of node 'j' that are the neighbours of node 'i'

This reduces the number of neighbouring nodes to be considered during the label update of a node. As well as, this reduces the number of iterations required for convergence. Moreover, since socially unrelated nodes are ignored during label update, the modularity(Q) measure of the determined communities increases. The entire steps are depicted in Algorithm 1.

Algorithm 1: Label Propagation

```
Initialize labels: for each v ∈ V, lv(0) = v;
 i=0;
 while the stop criterion is not met do
  i++;
  let π = (vπ(1), vπ(2), . . . , vπ(n)) be a random permutation of the
vertices.
  Propagation:
  for j = 1 to n do
```

$$l(i) = l(\underset{nb_u}{argmax}(score\_reply_{(nb_u)})) \tag{4}$$

$$score\_reply(nb_u) = \sum_{j \in nb}^{\square} (sim_{ij\square}.n(j)) \tag{5}$$

```
    end
   end
   return Final labeling: lv(t) for each v ∈ V , where t is the last executed
  step.
```

4 Evaluation

The performance of the Label propagation with modified update rule is incorporated into Twitter, a real world social network to detect concept based community. We analyze the quality of community detection using modularity measure and Normalized Mutual Information.

4.1 The Modularity Measure

The modularity Q is proposed by Newman and Girvan [18] as a measure of the quality of a particular division of a network, and is defined as follows:

Q = (number of edges within communities)-(expected number of such edges)

The basic idea is to compare the division to a null model", a randomized network with exactly the same vertices and same degree, in which edges are placed randomly without regard to community structure.

In other words, the modularity Q measures the fraction of the edges in the network that connect vertices of the same type, i.e., within-community edges, minus the expected value of the same quantity in a network with the same community division but with random connections between the vertices . If the number of within community edges is no better than random, Q = 0. Values of Q that are close to 1, which is the maximum, indicates strong community structure.Q typically falls in the range from 0.3 to 0.9 and high values are rare.

4.2 Normalized Mutual Information

In addition to modularity, we use normalized mutual information to evaluate the community division results. NMI is computed as follows:

$$NMI(X,Y) = \frac{I(X;Y)}{[H(X)+H(Y)]/2} \tag{6}$$

where I(X;Y) measures the mutual information and H is the entropy function

X={w1, w2, w3,... wk} is the set of clusters and Y ={c1, c2, c3,...cj} is the set of classes. We tested both the original Label update rule and modified Label update rule on the constructed concept graph using 100 runs. The average modularity value and NMI obtained during our test setup from extracted tweets is shown in Table 1.

Table 1. Evaluation of Detected Tweet Concept Community

	Average modularity	NMI
Original-LPA(standard label update rule)	0.767	0.396
Modified-LPA(modified label update rule)	0.851	0.451

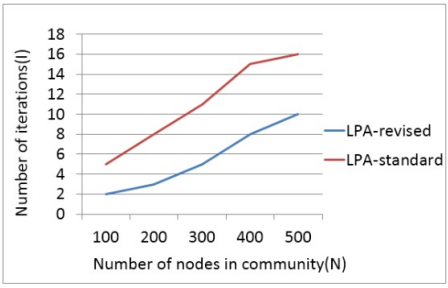


Fig. 2. Relation between the number of iterations for convergence to the size of nodes in each detected community

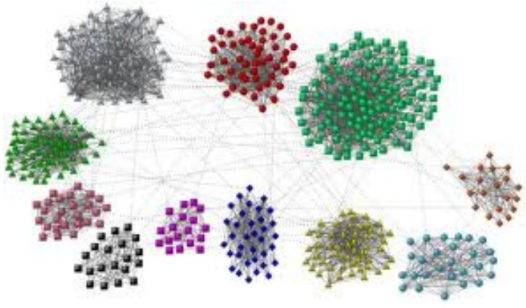


Fig. 3. Academic Concept communities detected using modified LPA

The convergent behaviour of the algorithm is shown in Fig 2. The figure shows that number of iterations grows logarithmically with the size of N, where N is the number of nodes in each of the detected community. The Fig.2 as well shows that, the number of iterations with modified update rule is less than standard update rule before convergence. Fig.3 shows the detected academic concept communities of the twitter network. The package igraph is used for visualization. In the network, the academic concept community like machine learning, computer network, software testing, data structure etc has been detected.

5 Conclusion

In this paper, we present a modified label propagation algorithm, which uses social relation in selecting the neighbours during label update. This improves the speed and quality of detected communities when compared to standard Label propagation algorithm. However, during label update, the first level of social relation alone is considered in deciding and adopting the new label. Our work can be extended by incorporating this approach. In our wok, pre-defined academic concepts are considered in detecting the communities. As an extension, instead of using pre-defined academic concepts, learning algorithm can be used to learn and classify the extracted tweet concept as an academic concept.

References

1. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks (WOSP 2008), pp. 19–24. ACM, New York (2008)
2. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, San Jose, California, August 12, pp. 56–65 (2007)
3. Honeycutt, C., Herring, S.C.: Beyond Microblogging: Conversation and Collaboration via Twitter. In: 42nd Hawaii International Conference on System Sciences HICSS, Big Island, HI, USA, pp. 1–10 (2009)
4. Antenucci, D., Handy, G., Modi, A., Tinkerhess, M.: Classification of Tweets via Clustering of Hashtags (2011), http://twiki.di.uniroma1.it/pub/ApprAuto/WebHome/AntenucciHandyModiTinkerhess.pdf
5. Naaman, M., Boase, J., Lai, C.-H.: Is it Really About Me? Message Content in Social Awareness Streams. In: CSCW 2010, Savannah, Georgia, USA, pp. 255–256 (2010)
6. Ritter, A., Cherry, C., Dolan, B.: Unsupervised modeling of Twitter conversations, Human Language Technologies. In: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, pp. 172–180 (2010)
7. Phuvipadawat, S., Murata, T.: Detecting a Multi-Level Content Similarity from Microblogs based on Community Structures and Named Entities. Journal of Emerging Technologies in Web Intelligence 3(1), 11–19 (2011)
8. Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical clustering of tweets. In: Proceedings of SIGIR Workshop on Social Web Search and Mining (2011)
9. Luiten, M., Kosters, W.A., Takes, F.W.: Topical Influence on Twitter: A Feature Construction Approach. In: Proceedings of 24th Benelux Conference on Artificial Intelligence(BNAIC 2012), pp. 139–146 (2012)

10. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76, 036106 (2007)
11. Zhu, X., Ghahraman, Z.: Learning from Labeled and Unlabeled Data with Label Propagation.Tech report CMU-CALD-02-107 (2002)
12. Šubelj, L., Bajec, M.: Unfolding network communities by combining defensive and offensive label propagation. In: Proceedings of the ECML PKDD Workshop on the Analysis of Complex Networks (ACNE 2010), Barcelona, Spain, pp. 87–104 (2010)
13. Xie, J., Szymanski, B.K.: Community detection using a neighborhood strength driven label propagation algorithm. In: IEEE NSW 2011, West Point, NY, pp. 188–195 (2011)
14. AlchemyAPI, http://www.alchemyapi.com
15. Twitter Streaming API, http://apiwiki.twitter.com/
16. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. Journal of Machine Learning Research 2, 419–444 (2002)
17. Martins, Mario A. T. Figueiredo, Pedro M. Q. Aguiar.: Kernels and similarity measures for text classification, In: 6th Conference on telecommunications—ConfTele 2007, Peniche, Portugal(2007)
18. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (2004)
19. Cordasco, G., Gargano, L.: Community Detection via Semi-Synchronous Label Propagation Algorithms, arxiv:1103.4550
20. Wang, F., Zhang, C.: Label Propagation through Linear Neighbourhoods. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 85–992. ACM New York (2006)
21. Chen, J., Ji, D., Tan, C.L., Niu, Z.: Relation extraction using label propagation based semi-supervised learning. In: ACL-44th Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, pp. 129–136 (2006)
22. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA 98(2), 404–409 (2001)
23. Fell, D.A., Wagner, A.: The small world of metabolism. Nature Biotechnology 18(11), 1121–1122 (2000)
24. Danon, L., Duch, J., Arenas, A., Diaz-guilera, A.: Comparing community structure identification. Journal of Statistical Mechanics:Theory and Experiment 9008, 09008 (2005)
25. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99(12), 7821–7826 (2002)
26. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70, 066111 (2004)
27. Barber, M.J.: Detecting network communities by propagating labels under constraints. Phys. Rev. E 80, 026129 (2009)
28. Leung, I.X.Y., Hui, P., Li, P., Crowcroft, J.: Towards real-time community detection in large networks. Phys. Rev. E 79, 066107 (2009)
29. Jia, G., Cai, Z., Musolesi, M., Wang, Y., Tennant, D.A., Weber, R.J.M., Heath, J.K., He, S.: Community Detection in Social and Biological Networks Using Differential Evolution. In: Hamadi, Y., Schoenauer, M. (eds.) LION 2012. LNCS, vol. 7219, pp. 71–85. Springer, Heidelberg (2012)
30. Bansal, S., Bhowmick, S., Paymal, P.: Fast community detection for dynamic complex networks. In: Mangioni, G. (ed.) CompleNet 2010. CCIS, vol. 116, pp. 196–207. Springer, Heidelberg (2011)

Automatic Tagging of Texts with Contextual Factors Using Knowledge Concepts

Rajendra Prasath[1], Philip O'Reilly[1], and Aidan Duane[2]

[1] University College Cork (UCC), Cork, Ireland
[2] Waterford Institute of Technology, Waterford, Ireland
{R.Prasath,Philip.OReilly}@ucc.ie, aduane@wit.ie

Abstract. We present a method to perform automatic tagging of contextual factors associated with mobile payments data. Users specify a short description about the contextual factors interesting to them. The proposed system characterizes these factors and generates the knowledge concepts, similar to [1,2], but with the help of corpus statistics. These knowledge concepts describe the factors in terms of multi-faceted information search. Secondly, given a query, the underlying retrieval system retrieves top k texts pertaining to user information needs. Then based on the similarity between each of the knowledge concepts and the best matching texts, the context matching score is computed. Then the ranked sequence of contextual tags are assigned to the each retrieved text. The experimental results show that the proposed approach characterizes the context from user specified factors and performs the contextual tagging of the retrieved texts in a better way.

Keywords: Knowledge Concepts, Contextual Tagging, Mobile Payments, Learning from Data.

1 Introduction

Web content is growing exponentially in various domains like finance, travel, health and so on. A vast amount of knowledge is hidden in these textual content. But, only a fraction of this knowledge is actually mined and utilized. Thus, the huge textual data is under-utilized, making it unable to support complex business decision making. Therefore, a significant opportunity exists to develop a method for mining contextual data from free text content. Pedersen and Kulkarni[3] illustrated the use of SenseClusters[1] in identifying similar contexts from natural language texts that may be a phrase/sentence/paragraph. Motivated by the distributional hypothesis [4]: "Texts with similar meaning tend to have similar contexts", in this paper, we propose a method to suggest the tags that mimics the context of the texts associated with mobile payment services.

By combining an organizations' internal knowledge with knowledge from external sources, as the knowledge concepts extracted from web sources in [1,2],

[1] http://senseclusters.sourceforge.net/

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 687–694, 2013.
© Springer International Publishing Switzerland 2013

organizations can achieve greater insight into business problems, opportunities, and indeed the entire market by capturing and learning similar contexts in a non-interactive way. This would therefore enable organizations to learn and apply similar contexts to address more effectively similar problems or opportunities, arising in the current and future business environments.

2 Mathematical Formulation

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of n contextual factors where each x_i, $1 \leq i \leq n$ represents a specific topic or tag closely associated with the domain D that is expressed in terms of a set of $|D|$ textual descriptions: $\{td_1, td_2, \cdots, td_{|D|}\}$, each td_j, $1 \leq j \leq N$ is assumed to represent an aspect associated with the domain. The choice of each x_i is selected by the users pertaining to their information needs. Each x_i should represent a meaningful *topic* or *aspect* of any information needs within the specific underlying domain D. Each contextual factor x_i can be expressed by a set of keywords / phrases $x_i = \{w_1, w_2, \cdots, w_m\}$ where each w_j, $1 \leq j \leq m$, may be either a keyword or a key phrase that is semantically associated with x_i. w_j is the part of the knowledge concepts $KC = \{kc_1, kc_2, \cdots, kc_k\}$ mined from the underlying web corpus. The extracted concept may be an illustration, description or explanation that characterize x_i. Now the task is to approximate the unknown target function $\phi : D \times X \rightarrow \{td_j, x_i\}$ by means of the function $\psi : KC \times X \rightarrow \{kc_k, x_i\}$ called a knowledge concept based classifier such that ϕ and ψ coincide as much as possible. In turn, the tag x_i is assigned to td_j when $sim(td_j, kc_k)$ is at its maximum with respect to x_i.

3 Automatic Tagging of Contextual Factors

The proposed approach performs automatic tagging of contextual factors based on the Pseudo Relevance Feedback (PRF) like approach with information fusion through knowledge concepts. At first, the proposed system characterizes the user specified contextual factors in the form of knowledge concepts that are extracted from web corpus.

Secondly, the proposed system retrieves top d documents pertaining to the information needs specified by the users. These retrieved texts are assumed to be (pseudo) relevant to the user information needs. Then we compute the context matching scores between each knowledge concept and each of the retrieved texts. Based on the scores, the ranked sequence of tags is assigned to the retrieved texts.

The proposed approach consists of two parts: *Building Knowledge Concepts* and *Contextual Tagging of Texts*. In the first part, the system performs understanding of user defined contexts with respect to their information needs and applies corpus statistics by means of co-occurrence information of user specified information needs. Using the corpus as the domain knowledge, top 5 textual descriptions are retrieved and augmented to form a knowledge concept pertaining to the user specified contextual factor. This process is repeated for

all contextual factors specified by the users. Each time, the system generates a knowledge concept for each contextual factor. The second part of the proposed approach is to perform automatic contextual tagging of texts using the knowledge concepts extracted for each of the user specified contextual factors. In this part, first we perform individual query term weighting to identify the focus of user information need. Query term weighting is computed as the multiplicative factor of the average term frequency and inverse document frequency. The term which gets higher weight is the most likely the focus of the user search. Next,

Algorithm 1. Building Knowledge Concepts

Require: Text Retrieval System
Input: A short description of the contextual factors

1: **Input:** User specifies the contextual factors of their information needs by a set of keywords
2: **for** each contextual factor T_i, $i \leq i \leq c$ **do**
3: **Corpus Statistics:** Using corpus statistics, span the contextual factors description by augmenting the co-occurring terms.
4: **Knowledge Concepts:** Use the augmented description to retrieve top 5 texts and incorporate them to form the knowledge concepts for T_i
5: **end for**

Output: A Knowledge Concept for each contextual factor

we build a topic model using Latent Dirichlet Allocation (LDA) on the entire corpus. Latent Dirichlet Allocation (LDA) is a generative probabilistic topic model [5,6]. To analyse a discrete collection of data, especially text corpora, LDA applies three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document. LDA defines the marginal distribution of a document as a continuous mixture distribution, as follows,

$$p_i(x) = p(d|\alpha_i, \beta_i) = \int p(\theta|\alpha) \left(\prod_{j=1}^{n} p(w_j|\theta, \beta) \right) d\theta \tag{1}$$

where d is a document with n words. $p(\theta|\alpha)$ and $p(w_j|\theta, \beta)$ are actually multinomial distributions with Dirichlet prior. $p(w_j|\theta, \beta)$ describes $K-$dimensional topic-word distributions. The parameters α, β are estimated by means of Gibbs sampling [7]. Here α is the symmetric Dirichlet prior for all documents and β is the symmetric Dirichlet prior for all topics.

Using the above topic model and the query term weighting, we perform topic assisted text retrieval in which texts covering many topics will be given higher weight. Then we compute the context matching scores of each of these extracted knowledge concepts with the retrieved texts and rank them based on the context matching scores. The selected set contextual factors which scores top l scores will be assigned as the tag of the retrieved text. The proposed procedure is illustrated in Algorithm. 2.

Algorithm 2. Automatic Tagging of Texts with Contexts

Require: Topic assisted Text Retrieval System
Input: Query Q having n keywords: $q_1, q_2, \cdots, q_n$

1: **Build Topic Models:** Use entire corpus to build the topic model using LDA
2: **Input:** Enter the user query into the system
3: **Query Terms Weighting:** Extract the query terms and use corpus statistics to determine their individual weights as follows:

$$score(q_i) = averageTF(q_i) * IDF(q_i), \forall i \in [1, n]$$

where $averageTF(q_i)$ is computed as the ratio between the total number of occurrences of the given query term across the textual descriptions in the collection and the total number of textual descriptions (in terms, the total number of textual descriptions in which the term occurs) and $IDF(q_i) = \log(\frac{N}{df(q_i)})$ where N = total number of textual descriptions and $df(q_i)$ = frequency of textual descriptions given the query term q_i.

4: **Topic Assisted Text Retrieval:** Compute $cosine(Q, T)$ with weighted query terms using cosine similarity and retrieve all texts; Based on the topic distributions, re-rank the retrieved texts.
5: **Cluster Similar Contexts**: Derive the Knowledge concepts for the chosen set of contextual factors; Compute and rank the context matching scores of each of these knowledge concepts with the retrieved texts.
6: **Assign Tags:** Based on the context matching scores, assign the ranked tag set to the retrieved texts.
7: **return** Texts with the ranked list of contextual tags

4 Experimental Settings

Web Corpus. We have created a collection of web documents crawled from websites of various financial companies using open source web crawler. We collected 14,764 web documents containing a total of 296,983 words. We selected the variations of 20 contextually different types of information pertaining to the mobile payments sector as seeds. We have used JGibbLDA[2] tool for building the topic models. We have selected 200 latent topics with 1000 iterations to build the topic model (assumed other LDA parameters: $\alpha = 0.5$ and $\beta = 0.1$).

The following user queries are used in our experiments: [Q1] What technology is needed to accept and process mobile payments? [Q2] Will mobile payments replace cash? [Q3] What are the factors which influence mobile payment adoption? [Q4] What are the benefits to consumers of mobile payments? [Q5] How secure are mobile payments?

4.1 Contextual Factors

In this experiment, we have considered 6 contextual factors each with its associated description. In the context of mobile payments, each factor can differ depending on the mobile payment platform, technology utilized and settings.

[2] `http://jgibblda.sourceforge.net`

Convenience (T1): The time and effort consumers expend in purchasing a product/service [8].

Ease of Use (T2): The degree to which a person believes that using a specific system is free from effort [9].

Peer Influence (T3): The influence that a peer group, observers, or an individual, exerts to encourage others to change their attitudes, values, and behaviours to conform to group norms [10].

Level of Trust (T4): Consumers understanding of the level of competence, benevolence, and integrity of the parties with which they are interacting [9]

NFC Technologies (T5): Near Field Communication (NFC) supported technologies for mobile payments [11].

Cost (T6): Cost (to customers / service providers) associated with mobile payments technologies

Ease of use has been an essential driver of consumer acceptance of mobile applications by users and to compete with the established mobile payments providers [12]. A user's feelings of trust toward an online service is another important determinant, especially in mobile payments [13]. O'Reilly *et* al. [11] believe that trust is the most important antecedent. Similarly, Mallat[14] found that trust in vendors and mobile network operators (MNOs) is essential to reduce consumers perceived risks of mobile payments. Convenience is also a key factor in determining whether mobile payment solutions are adopted[9]. Research has illustrated that a solution such as SquareUp, perceived by consumers as convenient in one specific context, can be considered less convenient in an alternative setting [11]. Peer influence has also been identified as a key factor in the adoption of new technologies including mobile phones [15]. Extant research has illustrated the role of peer influence in the adoption of mobile payments in different countries, with Asian countries having much higher adoption rates than Western countries [14]. NFC technologies have also been identified as a key enabler of mobile payments[9]. However NFC payment solutions are successful countries such as Hong Kong (Octupus card) with a 95% adoption rate [11].

4.2 Evaluation Methodology

We have performed the subjective evaluation to test the quality of the automatic context tagging:

- The focus is on evaluating the quality of the contextual tagging of retrieved texts using 2 evaluators to judge the quality of the tagging of top n texts
- Each evaluator used the 5-point scale [5-Very Good, 4-Good, 3-Moderate, 2-Poor, 1-Very Poor] to judge the quality of contextual tagging
- Evaluators judged the contexts of each query with retrieved texts on a 3-point scale [3-matching, 2-partial and 1-distinct contexts).

5 Results and Discussion

Table 1 presents a discussion and analysis of the data retrieved in terms of the specific user queries. Specifically, it presents an illustration of those pieces

of information which were ranked highest by the evaluators in terms of the contextual tagging (score of 5) and the text context (score of 3) and the proposed method for contextual tagging. Figure. 1 illustrates the performance of the contextual tagging across the five queries. Each of the 100 records retrieved were reviewed by three evaluators. The quality of the information relative to the query together with the sequence of the tagging was evaluated.

Table 1. Retrieved Texts with Suggested Tags

ID	Retrieved Texts
Q1	71% percent small businesses mobile technology accept mobile payments 52% utilize mobile- tablet-based point-of-sale system **Suggested Tags:** {T2, T5, T3, T4, T6, T1}
Q2	Easy alternative payments methods mobile square iZettle e-commerce cloud wallets approaching mass "replace" cards future **Suggested Tags:** {T3, T2, T1, T5, T4, T6}
Q3	Mechanisms driving adoption similar segments analysis revealed social networks social interactions influence mobile money uptake **Suggested Tags:** {T3, T4, T2}
Q4	Consumers benefit mobile payments convenience ability monitor finances control spending **Suggested Tags:** {T1,T4, T2, T3, T6, T5}
Q5	NFC offers greatest security combined biometrics fingerprint recognition makes phone completely secure mobile payments **Suggested Tags:** {T4, T6, T5, T3, T2, T1}

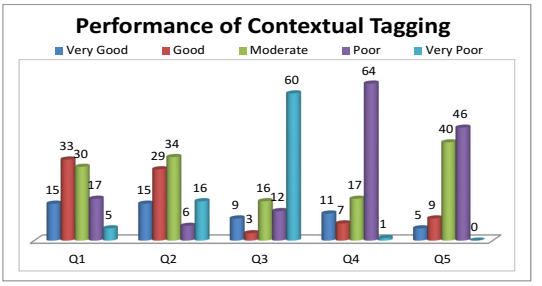


Fig. 1. Performance of Contextual Tagging for selected queries

The overall performance of tagging contextual factors is the good for queries 1 and 2. 48% of the tags retrieved were rated as good or better for query 1, with 44% being rated as good or better for query 2. Queries 3 and 4 performed poorest in terms of contextual tagging. Figure. 2 reveals that the overall quality of the information retrieved for these the selected queries in terms of the specific query being posed was relatively poor, hence providing an explanation for the high instances of distinct context. In terms of Q1, the excerpted query results (Table. 1) illustrate that convenience and ease of use were evident as the key

contextual factors pertaining to this query and Ease of Use is the key factor for small businesses to utilize mobile tablets at the Point of Sale. For mobile payments adoption, convenience and ease of use are key aspects. The ability of this method to abstract the association between acceptance of mobile payments and contextual factors such as convenience and ease of use is truly valuable. Q2 also performed to a high standard. The results illustrate that mobile payments may complement rather than replace cash. Key contextual factors such as trust and peer influence emerged as key indicators, which would be an accurate reflection of dedicated consumer surveys conducted to date on this subject.

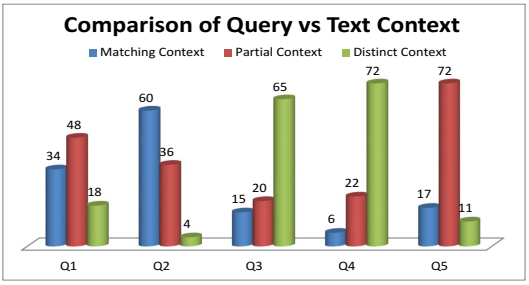


Fig. 2. Comparing the Query context with the retrieved text contexts

Q3 and Q4 performed poorly at a macro level (100 records) in terms of contextual tagging. In terms of determining why this is the case, analysis revealed a significant amount of noise in the data. The keyword adoption (Q3) and benefit (Q4) appears to be causing most of these are very generic terms across multiple disciplines. Nevertheless, a number of excellent snippets of information were retrieved (Table. 1). For Q3, the snippet illustrates that social networks are key in driving the adoption of mobile payments. Furthermore, the tagging correctly infers that peer influence is the key contextual factor. Even more valuable, the system correctly infers that social networks and social interactions are concepts associated with (aspects of) peer influence. The ability of the algorithm to reveal such associations is excellent. Similar observations are true for Q4 and Q5. Our observations with top 100 texts illustrate the efficiency of the proposed algorithm in terms of its ability to retrieve relevant information to the query and the ability to identify valuable contextual factors. This has significant value in order to provide valuable inferences for the user.

6 Conclusion

In this work, we presented a method to perform automatic tagging of contextual factors associated with mobile payments data. Users specify a short description about the contextual factors interesting to them. The proposed system characterizes the these factors and generates the knowledge concepts

using corpus statistics. These knowledge concepts describe the factors in terms of the fusion of multi-faceted information. Secondly, given a query, the underlying retrieval system retrieves top k texts pertaining to user information needs. Then based on the context score between each of the knowledge concepts and the retrieved texts, the ranked sequence of contextual tags are assigned to the each retrieved text. The experimental results show that the proposed approach characterizes the context from user specified factors and performs the contextual tagging of the retrieved texts in a better way.

References

1. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proc. of the 20th Int. Joint conf. on Artifical intelligence, IJCAI 2007, pp. 1606–1611. Morgan Kaufmann Publishers Inc. (2007)
2. Prasath, R., Sarkar, S.: Unsupervised feature generation using knowledge repositories for effective text categorization. In: Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence, pp. 1101–1102. IOS Press, Amsterdam (2010)
3. Pedersen, T., Kulkarni, A.: Identifying similar words and contexts in natural language with senseclusters. In: Proc. of the 20th National Conf. on Artificial Intelligence, AAAI 2005, pp. 1694–1695. AAAI Press (2005)
4. Johnson, W., Lindenstrauss, L.: Extensions of lipschitz maps into a hilbert space. Contemporary Mathematics 26, 189–206 (1984)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
6. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (2012)
7. Griffiths, T.: Gibbs sampling in the generative model of latent dirichlet allocation. Technical report, Stanford University (2002)
8. Brown, L.G.: Convenience in services marketing. Journal of Services Marketing 4(1), 53–59 (1990)
9. Duane, A., O'Reilly, P., Andreev, P.: Trusting m-payments - realising the potential of smart phones for m-commerce: A conceptual model & survey of consumers in ireland. In: ICIS 2011 (2011)
10. Kelman, H.C.: Compliance, identification, and internalization three processes of attitude change. Journal of Conflict Resolution 2(1), 51–60 (1958)
11. O'Reilly, P., Duane, A., Andreev, P.: To m-pay or not to m-pay - realising the potential of smart phones: conceptual modeling and empirical validation. Electronic Markets 22(4), 229–241 (2012)
12. Schierz, P.G., Schilke, O., Wirtz, B.W.: Understanding consumer acceptance of mobile payment services: An empirical analysis. Electron. Commer. Rec. Appl. 9(3), 209–216 (2010)
13. Roca, J.C., García, J.J., de la Vega, J.J.: The importance of perceived trust, security and privacy in online trading systems. Inf. Manag. Comput. Security 17(2), 96–113 (2009)
14. Mallat, N.: Exploring consumer adoption of mobile payments - a qualitative study. J. Strateg. Inf. Syst. 16(4), 413–432 (2007)
15. Lee, S.Y.: Examining the factors that influence early adopters' smartphone adoption: The case of college students. Telematics and Informatics (to appear, 2013)

MetaProPOS++: An Automatic Approach for a Meta Process Patterns' Ontology Population

Nahla Jlaiel[1], Refka Aissa[2], and Mohamed Ben Ahmed[1]

[1] ENSI, National School of Computer Science,
University of La Manouba, La Manouba 2010, Tunisia
{Nahla.Jlaiel,Mohamed.Benahmed}@riadi.rnu.tn
[2] ESCT, Higher School of Business,
University of La Manouba, La Manouba 2010, Tunisia
Aissa_Refka@hotmail.fr

Abstract. This paper deals with ontology population in the context of building a semantic framework for software process patterns capitalization and reuse improvement. In this paper, we propose an automatic approach for an existing ontology population, named MetaProPOS. This ontology aims to unify different and heterogeneous software patterns' descriptions coming from diverse patterns' collections (e .g. Ambler, Störrle, Gamma, Coulette, Conte, Ribo, etc.). This paper provides also a survey of ontology population approaches and systems. This survey serves as a basis for the choices we made in order to set up the proposed approach MetaProPOS++. In addition, a description and an empirical evaluation of the implemented solution, MetaProPOP, is presented in this paper, giving more details on our proposition.

Keywords: Semantic Web, Ontology population, Software patterns, XML, OWL, RDF, Jena, SPARQL, Triple stores, MetaProPOS, MetaProPOS++, MetaProPOP.

1 Introduction

Within the software engineering communities, patterns have been increasingly recognized as an effective method to reuse knowledge and best practices gained during the whole software lifecycle. Thus, software patterns now exist for a wide range of topics including requirement, analysis, design, implementation or code patterns, test patterns and even maintenance patterns. Concerning process patterns, they are widely used by the software development community as an excellent medium to share software development knowledge that is often encapsulated in experiences and best practices. In other words, they capitalize good specifications or implementations of a method to be followed to achieve a goal [1]. As consequence to the huge proliferation of the process patterns practice, several description models and languages were proposed and used in software development research and practice including AMBLER [2], RHODES [3], GNATZ [4], P-SIGMA [5], STÖRRLE [6], PROMENADE [7], PPDL [8], PROPEL [9], PLMLx [10], UML-PP [11] and PPL [12], to name just a few.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 695–708, 2013.
© Springer International Publishing Switzerland 2013

These latter suffer from serious problems related lacks of patterns' formalization, patterns' unification and patterns' usage assistance and guidance that we developed in a previous work [1]. As a matter of fact, process patterns are often being used in an informal manner, through traditional textbooks or better with modest hypertext systems providing weak semantic relationships. In addition to the huge number of process patterns that are available in books or Web-based resources, they significantly differ in format, coverage, scope, architecture and terminology used.

Since the overall objective of our research is to build up an intelligent framework in order to ease process patterns capitalization and reuse [13], a first step in this direction was to represent different process patterns in a unified and formal manner in order to allow rigorous reasoning process on process patterns' knowledge whatever the format and the terminology used are. To do so, we proposed a meta process patterns ontology, named MetaProPOS [1], providing common and shared architecture, terminology and semantics for patterns' unification, mediation and also mining [14].

In this paper, we propose an automatic approach for the MetaProPOS ontology population, named MetaProPOS++. This latter is performed using the Jena semantic Web framework [15] and implementing two different population techniques.

So, the remainder of this paper is organized as follows: section 2 provides background information on first, the overall context of this work and second, on the aimed ontology MetaProPOS. Section 3 summarizes the study that we carried out on ontology population approaches and systems. Section 4 details the proposed approach MetaProPOS++ by describing the two proposed techniques for ontology population. Section 5 provides an empirical evaluation of the implemented system called MetaProPOP by comparing the two techniques and presenting a synthesis of the evaluation results. Section 6 concludes the paper by giving a discussion of our contribution and some future directions.

2 Background

In this section, we present the context and the motivation of the current work by giving background information on the overall approach and the targeted ontology.

2.1 A Semantic Approach for Patterns' Capitalization and Reuse

In order to enhance the quality and capacity of software patterns' capitalization and reuse, a first step in this direction was to build a unified and shared conceptualization of patterns through the proposed ontology MetaProPOS presented in the following subsection. A subsequent initiative is to show how the ontology is performed to achieve the aforementioned goals through the SCATTER approach [14]. Acronym for "SemantiC Approach for sofTware process patTErns capitalization and Reuse" and as

its name implies, this approach aims to improve the quality and capacity of patterns' sharing through formal and semantic technique of knowledge capitalization and reuse. SCATTER is based on two main processes:

Triple Unification. Given different collections of software patterns, this process consists of a triple unification effort addressing three different levels (Terminological, Architectural and Semantic) allowing automatic and implicit conversion of heterogeneous and unstructured software patterns to structured, unified and semantic ones. This is ensured by means of a natural language processing technique coupled with a population ontology approach using terminological (WordNet, Wolf) as well as ontological (MetaProPOS) resources.

Patterns' Mining. It consists of a reasoning process carried out on the stored patterns forming a knowledge base which is exploited by an inference engine relying on the ontology proposed for this purpose, MetaProPOS and a system of logical rules based on Prolog. Indeed, the patterns' mining process would create a real algebra of patterns allowing a better search of candidate patterns regarding to a context and a given problem for instance, by exploring the patterns' relationships facet (similarity, use, refinement, alternative, etc.)

2.2 A Meta Process Patterns' Ontology for Software Development

Acronym for Meta Process Patterns' Ontology for Software development, MetaProPOS [1] consists of a formal and unified representation of software process patterns providing architectural and semantic unification of patterns knowledge in addition to an Inferential basis for building an intelligent framework of process patterns' capitalization and reuse within software development communities. Indeed, the proposed ontology aims to interconnect different process patterns' collections whatever the pattern's format is or the terminology employed is in order to uniformly express and share patterns knowledge.

To do so, we built an OWL ontology based on Description Logics (OWL-DL) to unify software patterns descriptions according to eleven facets (cf. Fig. 1). The identification facet encapsulates a set of properties identifying a pattern such as pattern name, author(s), keywords, pattern's classification (type, category, abstraction level, and aspect) as well as pattern origin (project and participants) and pattern artifacts (used and/or produced). The core information is the main pattern facet embodying details about the well-known triplet: problem, context and solution. The relationships facet expresses how a pattern could interact with other patterns (e.g. similar patterns, refinement patterns, subsequent patterns, and anti-patterns). The guidance facet refers to the support level provided by a pattern to be comprehended and used (e.g. known uses, example, literature, illustration, etc.). The evaluation facet provides feedbacks on pattern application (e.g. discussion, confidence, maturity, etc.). The management facet provides general information about a given pattern (e.g. version, creation-date).

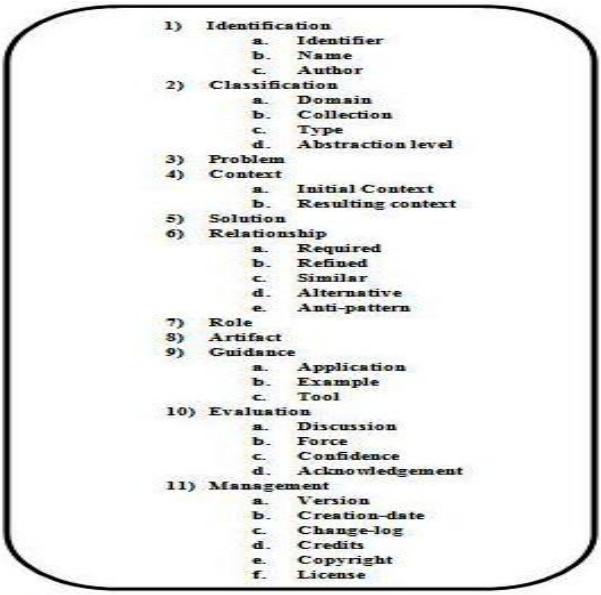


Fig. 1. The adopted unified description of software patterns

3 Related Work

In order to populate MetaProPOS, we conducted a literature review focusing on ontology population that revealed different approaches and systems. These works are classified into three main categories, namely: Automatic Ontology Population, Semi-Automatic Ontology Population and Flexible Ontology Population.

3.1 Automatic Ontology Population

This category covers four main works. The proposed method in [16] is based on an unsupervised technique considering that if a term T belongs to a class C, then in a text corpus we may expect the occurrence of phrases like such C as T.

The presented work in [17] proposes an unsupervised technique for ontology population based on vector-feature similarity between each concept C and a term to be classified T, considering the word of the class as a pivot word for acquiring relevant contexts for the class.

The Class-Example approach described in [18] considers the similarity between a term T and a set of training examples which represent a certain class. This system outperforms significantly the other two methods, making it appealing even considering the need

of supervision. In [19] they describe an approach using instances of some classes returned by Google search to find instances of other classes, populating then a movie's domain ontology.

3.2 Semi-automatic Ontology Population

This category concerns two main approaches. The first one described in [20], details the process of ontology pre-population based on a supervised machine learning technique.

The second approach presented in [21], describes how to populate ontologies from textual documents using an environment for mapping the linguistic extractions with the domain ontology based on the use of knowledge acquisition rules.

3.3 Flexible Ontology Population

This category includes two main works. The first one described in [22], proposes a GATE resource called the OwlExporter that allows easily mapping existing NLP analysis pipelines to OWL ontologies by allowing language engineers to create ontology population systems without requiring extensive knowledge of ontologies management systems.

The second proposition [23], presents a flexible method consisting in populating an existing OWL ontology from XML data. This method is based on the definition of a rules' graph representing the mapping from XSD schema elements to OWL schema ones. Table 1 summarizes the studied approaches for ontology population where (--) means unsupported criteria.

Table 1. Summary table of the ontology population approaches

Approach	Population type	Learning type	Source type	Tool	License
[16]	Automatic	Unsupervised	Text	--	--
[20]	Semi-Automatic	Supervised	Text	Ontosophie	Proprietary
[17]	Automatic	Unsupervised	Text	--	--
[21]	Semi-Automatic	--	Text	OntoPop	Proprietary
[18]	Automatic	--	Text	--	--
[19]	Automatic	--	Text	--	--
[22]	Flexible	--	Text	OWLExporter	Open source
[23]	Flexible	--	XML	--	--

In addition to ontology population approaches, we also studied ontology population systems including WEB→KB [23], BOEMIE [24], Adaptiva [25], KnownItAll [26], ArtEquaAKT [27], SOBA [28], LEILA [29], ISOLDE [30], OPTIMA [31], etc. that are summarized in Table 2.

Table 2. Summary table of ontology population systems

System	Creation Date	Source type	Accessibility	Used technique
WEB→KB	1997	Web documents	--	Logics
BOEMIE	2000	Text corpora	--	Machine learning
Adaptiva	2002	Text documents	--	Bootstrapping
KnownItAll	2004	Unstructured texts	--	Statistics
SOBA	2006	Football matches reports	--	Rules
ArtEquAKT	2004	Artistic bibliographies	Demo	Heuristics
LEILA	2006	Text corpora	--	Statistics
ISOLDE	2006	Ontologies' seeds	--	Statistics
OPTIMA	2008	Structured/semi-structured texts	--	Linguistic

4 MetaProPOS++

In order to populate MetaProPOS, we choose to create a custom approach specific to the application domain of software patterns.

To do so, we first, went through a process of Information Extraction using a Natural Language Processing technique converting heterogeneous descriptions of patterns into a XML format corresponding to the unified proposed description [32].

A second process consists of transforming XML patterns to MetaProPOS OWL instances. This is performed via the proposed approach MetaProPOS++ and using the Jena OWL API (cf. Fig. 2). Indeed, to achieve this goal, we adopted two population techniques, memory-based and triple store-based techniques that are implemented using Jena.

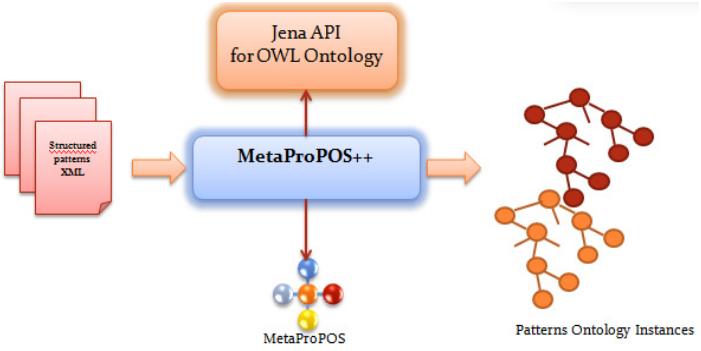


Fig. 2. Overview of the proposed approach

4.1 Memory-Based Population

This paragraph presents an automatic and semantic technique, called In-memory ontology population ensuring individuals conversion from XML patterns to OWL patterns. The idea is to create a in-memory RDF graph representing the ontological model, in order to maintain the population that allowing the generation of a novel URI for each novel instance. This URI represents a novel instance to be inserted in the MetaProPOS ontology which is composed of three elements (subject, predicate and literal) called triple. After population, the model will be serialized in the RDF/XML format, Turtle format or N-Triple format in order to make information persistent (cf. Fig. 3).

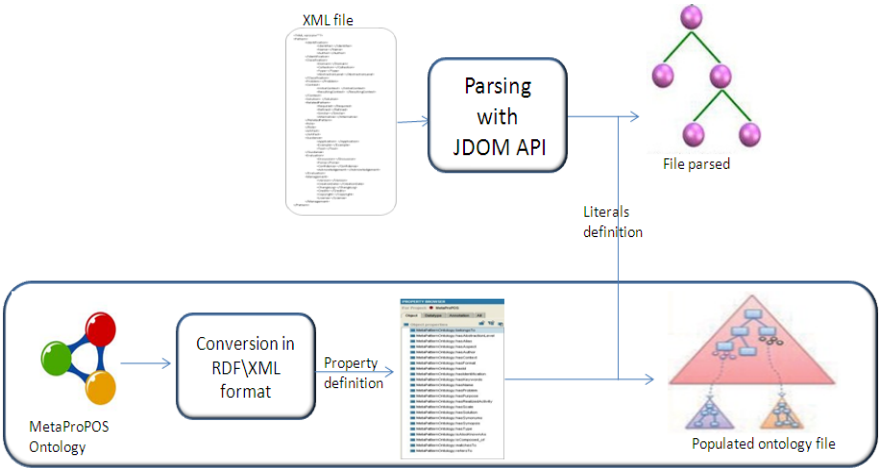


Fig. 3. An overview of the in-memory ontology population technique

From the information kept after the parsing, the system would be ready to establish the relation between each literal and each property, and in this way all the triples are integrated into the MetaProPOS ontology.

4.2 Triple Store-Based Population

This technique uses an RDF triple store to populate the OWL ontology implying a separation between the ontology and the patterns instances. Fig. 4 illustrates the functioning of the second technique, called triple-store population technique in which we use an RDF server that acts as a linker between the triple store and the SPARQL query to populate instances into the MetaProPOS ontology.

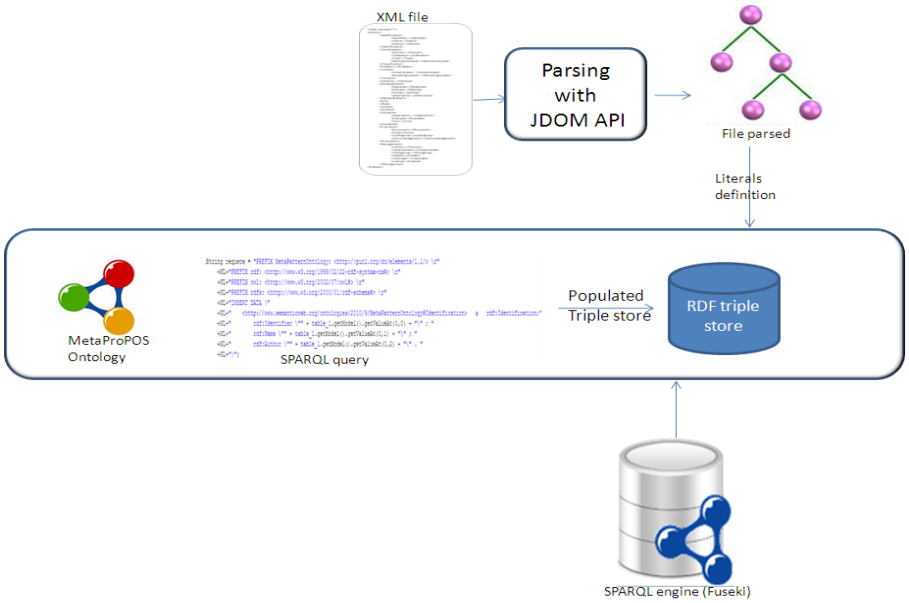


Fig. 4. An overview of the triple-store ontology population technique

5 Empirical Evaluation of MetaProPOP

So as to evaluate the MetaProPOP system, which is the implementation of our proposed approach MetaProPOS++, we experimented the approach and observed the system's behavior in terms of *occupied size* as well as *response time*. These experimentations were made on seven available training patterns.

The size occupied is an assessment of the size consumed during the population of instances in the MetaProPOS ontology and is expressed in Kilo Bytes. The response time, expressed in milliseconds, is the time elapsed between the launch of the population until its endings. The remainder of this section describes the two experimentation results of the two implemented techniques.

5.1 Experimentation Results of the In-Memory Population Technique

This subsection reports the empirical behavior of the MetaProPOP system when using in-memory population technique. The system performance is evaluated in terms of response time and occupied size.

Table 3 shows the occupied sizes variability with respect to different patterns sizes according to the in-memory population technique. We note that the in-memory population technique is not expensive according to the size of patterns being integrated.

Table 3. Evaluation results of the in-memory technique in term of size (in Kilo Bytes)

Patterns Number	0	1	3	7
Pattern Size	--	5	13	20
Occupied Size	86	176	182	191

Table 4 confronts response times when selecting a pattern to be in-memory populated throughout four iterations. We note that the response time is on average around 5.0E-4ms after four iterations.

Table 4. Evaluation results of the in-memory population technique in term of response time by a uni-selection of pattern

XML Pattern	Size	Iter1	Iter2	Iter3	Iter4	Average
Business	5 kb	7.0E-4ms	3.0E-4ms	5.0E-4ms	4.0E-4ms	5.0E-4ms
Specification	6 kb	5.0E-4ms	4.0E-4ms	3.0E-4ms	4.0E-4ms	4.0E-4ms
Design	7 kb	3.0E-4ms	4.0E-4ms	3.0E-4ms	3.0E-4ms	3.0E-4ms
Analysis	5 kb	4.0E-4ms	5.0E-4ms	3.0E-4ms	6.0E-4ms	5.0E-4ms
Review	6 kb	3.0E-4ms	5.0E-4ms	4.0E-4ms	3.0E-4ms	4.0E-4ms
Diverse	3kb	3.0E-4ms	4.0E-4ms	3.0E-4ms	5.0E-4ms	4.0E-4ms
Architecture	3kb	3.0E-4ms	4.0E-4ms	3.0E-4ms	5.0E-4ms	4.0E-4ms

Table 5 shows that the response size is, on average, around 3.0E-4ms for the in-memory population of the 7 considered patterns.

Table 5. Evaluation results of the in-memory population technique in term of response time by a multi-selection of patterns

	Patterns Number	Patterns total Size	Iter1	Iter2	Iter3	Iter4	Average
Patterns	7	38kb	5.7E-4ms	2.6E-4ms	1.8E-4ms	3.1E-4ms	3.0E-4ms

5.2 Experimentation Results of the Triple-Store Population Technique

This subsection reports the empirical behavior of the MetaProPOP system when using the triple-store population technique. This technique uses an RDF store and does not keep a large graph in memory unlike the in-memory. The MetaProPOP system performance is evaluated also, in terms of response time and occupied size. Table 6 shows the occupied sizes variability with respect to different patterns sizes according to the triple store population technique. We note that the triple-store population technique requires much space according to the size of patterns being integrated.

Table 6. Evaluation results of the triple-store technique in term of size (in Kilo Bytes)

Patterns Number	0	1	3	7
Pattern Size	--	5	13	20
Occupied Size	1.85	388	393	416

Table 7 shows that the response time is higher taking, on average, around 6.5E-4ms for the triple-store population of the 7 considered patterns.

Table 7. Evaluation results of the triple-store population technique in term of response time by a uni-selection of pattern

XML Pattern	Size	Iter1	Iter2	Iter3	Iter4	Average
Business	5 kb	8.0E-4ms	5.0E-4ms	5.0E-4ms	7.0E-4ms	6.0E-4ms
Specification	6 kb	8.0E-4ms	7.0E-4ms	6.0E-4ms	5.0E-4ms	7.0E-4ms
Design	7 kb	9.0E-4ms	7.0E-4ms	5.0E-4ms	5.0E-4ms	8.0E-4ms
Analysis	5 kb	8.0E-4ms	7.0E-4ms	5.0E-4ms	5.0E-4ms	6.0E-4ms
Review	6 kb	8.0E-4ms	8.0E-4ms	5.0E-4ms	6.0E-4ms	7.0E-4ms
Diverse	6 kb	7.0E-4ms	7.0E-4ms	6.0E-4ms	5.0E-4ms	6.0E-4ms
Architecture	3 kb	8.0E-4ms	7.0E-4ms	6.0E-4ms	6.0E-4ms	7.0E-4ms

Table 8 shows that the response size is more interesting consuming, on average, around 1.61E-4ms for the triple store population of the 7 considered patterns.

Table 8. Evaluation results of the triple-store population technique in term of response time by a multi-selection of patterns

	Patterns Number	Patterns total Size	Iter1	Iter2	Iter3	Iter4	Average
Patterns	7	38 kb	1.66E-4ms	1.59E-4ms	1.56E-4ms	1.62E-4ms	1.61E-4ms

5.3 Synthesis

Given these results and observations, we should ask what would be the best semantic and automatic ontology population technique. This subsection presents the synthesis of the empirical evaluation of the MetaProPOP system while answering this question. Indeed, Figure 5 shows that the triple-store ontology population technique is more space-consuming (416 kb) than the in-memory technique (191 kb).

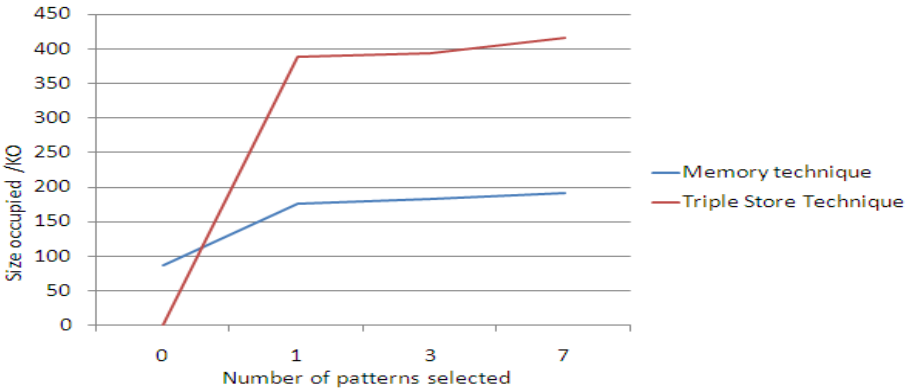


Fig. 5. Comparison of the sizes variance in MetaProPOP

In addition, the histogram of the Figure 6 shows that the in-memory technique is faster if we consider only one pattern for ontology population.

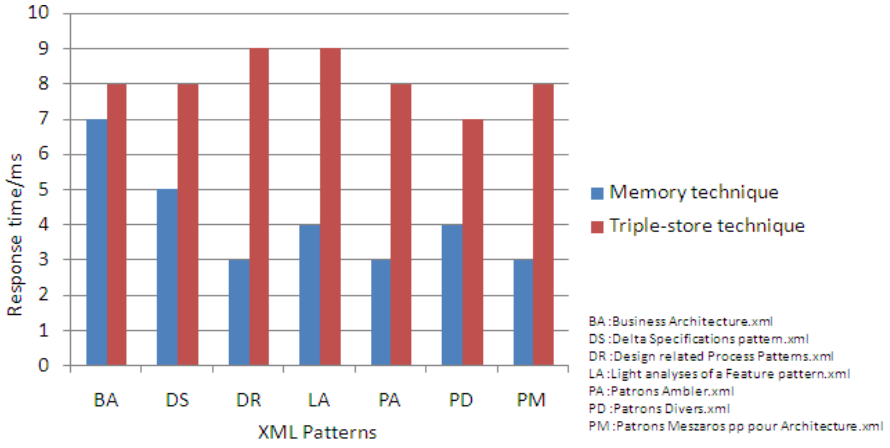


Fig. 6. Comparison of the MetaProPOP response times variance for one selected pattern

However, the confrontation of the results of the two methods of triple-store ontology population technique (uni-selection and multi-selection) shows that the technique based on triple store by multi-selections of patterns (1.61E-4ms) is faster than the Uni-selection patterns one (47.0E-4ms). Indeed, we note through Fig. 7, that in the first iteration the MetaProPOP system is much faster using the triple-store technique then using the in-memory technique.

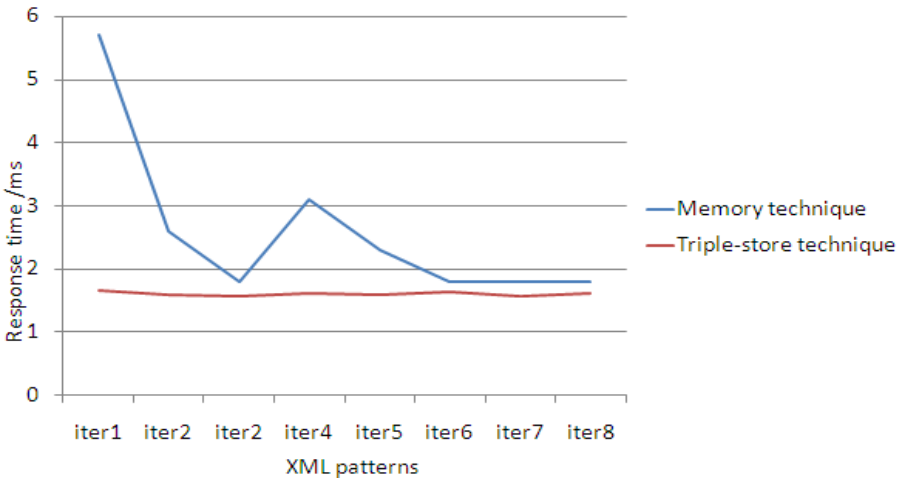


Fig. 7. Comparison of the MetaProPOP response times variance for all selected patterns

As a conclusion, we argue that the MetaProPOP system would be able to automatically and implicitly choose between the two implemented techniques the fasted one according to the population context and needs.

6 Conclusion and Future Work

The main contribution of this paper is the proposed semantic approach for the automatic MetaProPOS population, named MetaProPOS++. This latter uses the OWL Jena API to manage the MetaProPOS ontology for a unified software patterns description. An another contribution of this work is the MetaProPOP system which implements the proposed approach using two different techniques of ontology population based on Jena, namely: in-memory and triple-store based population techniques. The experimental evaluation of MetaProPOP shows that in-memory technique (191 Kb) is better than the triple store technique according to the size occupied (416 Kb). However, it shows that the triple-store technique is faster when multiple patterns descriptions are selected. The proposed approach MetaProPOS++, provides a good starting point as well as a strong foundation for capitalization and reuse of unified and semantic annotated patterns.

Several improvements for the MetaProPOP system are possible. As a first one, we are working on the integration of an inference engine to exploit the knowledge base created from the population process using a system of logics rules.

References

1. Jlaiel, N., Ben Ahmed, M.: MetaProPOS: a meta-process patterns ontology for software development communities. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS, vol. 6881, pp. 516–527. Springer, Heidelberg (2011)

2. Ambler, S.W.: Process Patterns: Building Large-Scale Systems Using Object Technology. Cambridge University Press/SIGS Books, Cambridge (1998)
3. Coulette, B., Crégut, X., Dong, T.B., Tran, D.T.: RHODES, a Process Component Centered Software Engineering Environment. In: The Proceedings of the 2nd International Conference on Enterprise Information Systems, Stafford, pp. 253–260 (2000)
4. Gnatz, M., Marschall, F., Popp, G., Rausch, A., Schwerin, W.: Towards a Living Software Development Process Based on Process Patterns. In: Ambriola, V. (ed.) EWSPT 2001. LNCS, vol. 2077, pp. 182–202. Springer, Heidelberg (2001)
5. Conte, A., Fredj, M., Giraudin, J.P., Rieu, D.: P-Sigma: A Formalism for A Unified Representation of Patterns (in French). In: 19ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision, Martigny , pp. 67–86 (2001)
6. Störrle, H.: Describing Process Patterns with UML. In: Ambriola, V. (ed.) EWSPT 2001. LNCS, vol. 2077, pp. 173–181. Springer, Heidelberg (2001)
7. Ribó, J.M., Franch, X.: Supporting Process Reuse in PROMENADE, Research report, Politechnical University of Catalonia (2002)
8. Dittmann, T., Gruhn, V., Hagen, M.: Improved Support for the Description and Usage of Process Patterns. In: The 1st Workshop on Process Patterns, 17th ACM Conference on Object-Oriented Programming, Systems, Languages and Applications, Seattle, pp. 37–48 (2002)
9. Hagen, M., Gruhn, V.: Towards Flexible Software Processes by using Process Patterns. In: The 3rd IASTED Conference on Software Engineering and Applications, Cambridge, pp. 436–441 (2004)
10. PLMLx, http://www.cs.kent.ac.uk/people/staff/saf/patterns/diethelm/plmlx_doc
11. Tran, H.N., Coulette, B., Dong, B.T.: Modeling Process Patterns and Their Application. In: The 2nd International Conference on Software Engineering Advances, Cap Esterel. IEEE Proceedings, pp. 15–20 (2007)
12. Meng, X.X., Wang, Y.S., Shi, L., Wang, F.J.: A Process Pattern Language for Agile Methods. In: The 14th Asia-Pacific Software Engineering Conference, Nagoya, pp. 374–381 (2007)
13. Jlaiel, N., Ben Ahmed, M.: Reflections on How to Improve Software Process Patterns Capitalization and Reuse. In: 9th International Conference on Information and Knowledge Engineering, pp. 30–35. CSREA Press, Las Vegas Nevada (2010)
14. Jlaiel, N., Ben Ahmed, M.: Towards a Novel Semantic Approach for Process Patterns Capitalization and Reuse. In: Proceedings of the 24th International Conference on Software Engineering & Knowledge Engineering (SEKE 2012), San Francisco Bay, USA, pp. 505–510 (2012)
15. http://incubator.apache.org/jena/
16. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, NJ, USA, pp. 539–545 (1992)
17. Cimiano, P., Völker, J.: Towards Large-scale, Open-domain and Ontology-based Named Entity Classification. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Bulgaria, pp. 166–172 (2005)
18. Tanev, H., Magnini, B.: Weakly Supervised Approaches for Ontology Population. In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge, vol. 167. IOS Press (2008)
19. Geleijnse, G., Korst, J.: Automatic Ontology Population by Googling. In: Proceedings of the 17th Conference on Artificial Intelligence, Belgium, Netherlands, pp. 120–126 (2005)

20. Celjuska, D., Vargas-Vera, M.: Ontosophie: A Semi-Automatic System for Ontology Population from Text. In: Proceeding of International Conference on Natural Language Processing ICON, India (2004)
21. Amardeilh, F., Laublet, P., Minel, J.-L.: Document Annotation and Ontology Population from Linguistic Extractions. In: Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP), Alberta, Canada, pp. 161–168 (2005)
22. Witte, R., Khamis, N., Rilling, J.: Flexible Ontology Population from Text: The OwlExporter. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, pp. 3845–3850 (2010)
23. Cruz, C., Christophe, N.: A Graph-based Tool for the Translation of XML Data to OWL-DL Ontologies. In: Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD), Paris, France, pp. 361–364 (2011)
24. http://www.webkb.org/
25. Brewster, C., Ciravegna, F., Wilks, Y.: User-Centred Ontology Learning for Knowledge Management. In: Andersson, B., Bergholtz, M., Johannesson, P. (eds.) NLDB 2002. LNCS, vol. 2553, pp. 203–207. Springer, Heidelberg (2002)
26. Etzioni, O., Kok, S., Soderland, S., Cagarella, M., Popescu, A.M., Weld, D.S., Downey, S.T., Yates, A.: Web-Scale Information Extraction in KnowItAll (Preliminary Results). In: Proceedings of the 13th International World Wide Web Conference (WWW 2004), New York, pp. 100–110 (2004)
27. http://www.iam.ecs.soton.ac.uk/projects/463.html
28. Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M.: Ontology-based Information Extraction with SOBA. In: Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy, pp. 2321–2324 (2006)
29. Suchanek, F.M., Ifrim, G., Weikum, G.: LEILA: Learning to Extract Information by Linguistic Analysis. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, OLP 2006, Sydney, Australia, pp. 18–25 (2006)
30. Buitelaar, P., Weber, N., Cimiano, P.: Ontology Learning and Population in SmartWeb. In: Proceeding of the Philips Symposium on Intelligent Algorithms (SOIA), Netherlands (2006)
31. Sang-Soo, K., Jeong-Woo, S., Seong-Bae, P., Changki Lee, J.H., Myung-Gil, J., Hyung-Geun, P.: OPTIMA: An Ontology Population System. In: Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3), Patras, Greece (2008)
32. Jlaiel, N., Madhbouh, K., Ben Ahmed, M.: A Semantic Approach for Automatic Structuring and Analysis of Software Process Patterns. The International Journal of Computer Applications 54(15), 24–31 (2012)

Discovery of Common Nominal Facts for Coreference Resolution: Proof of Concept*

Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences

Abstract. This paper reports on the preliminary experiment aimed at verification whether extraction of nominal facts corresponding to world knowledge from both structured and unstructured data could be effectively performed and its results used as a source of pragmatic knowledge for coreference resolution in Polish. Being the proof-of-concept only, this approach is work in progress and is intended to be further validated in a full-scale project.

1 Introduction

Coreference resolution is traditionally defined as a process of determining which fragments of a text correspond to the same discourse-world entities. As such, it is usually performed in two steps:

1. identifying *mentions* (or *markables*), i.e. phrases denoting entities in question
2. clustering mentions which denote the same referent.

The current scope of interest in research on coreference resolution for Polish is direct nominal coreference, i.e. identity-of-reference (in contrast to other anaphoric phenomena such as identity-of-sense anaphora, ellipsis, bound anaphora or bridging anaphora), with mentions being nominal groups (including single nouns, pronouns etc.). Following this assumption, the Polish Coreference Corpus [1] and coreference resolution tools [2,3] have been created, offering possibility to continue research on the subject.

The state-of-the art coreference resolution tools for Polish employ four extensive groups of features:

1. surface features (e.g. linking orthographic entity name with its abbreviation)
2. syntactic features (e.g. traditional gender/number agreement)
3. semantic features (e.g. agreement between semantic classes of mention heads)
4. discourse features (e.g. salience of topics).

Such approach results in a sufficiently effective (as compared to other languages) resolution process, but analysis of remaining errors reveals its one shortage: lack of representation of the world knowledge leads to clustering misses, affecting the final score of the whole process. Introducing pragmatic features representing widely known facts would, as we believe, increase probability of linking mentions denoting the same discourse-world object. In this paper we intend to verify whether this assumption is true before it can be applied in a large-scale project.

* The work reported here was carried out within the *Computer-based methods for coreference resolution in Polish texts (CORE)* project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40).

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 709–716, 2013.
© Springer International Publishing Switzerland 2013

2 Analysis of the Problem: Going Beyond Semantics

Currently available semantic bases such as WordNet (and its Polish equivalents: plWordNet [4] and POLNET [5]) only partially resolve this issue, by offering coarse semantic classes and semantic network traversal with via hypo-/hyperonymy/synonymy relations, which is not sufficient for texts abundantly using semantically close nominal phrases, making their automatic clustering problematic.

At the same time, it very often happens that there is no semantic connection between coreferent phrases whatsoever. In the following example:

(1) *Aldrin and Armstrong przyjaźnili się nadal, mimo że cała uwaga mediów skupiła się wyłącznie na pierwszym człowieku na Księżycu.*
'Aldrin and Armstrong stayed friends even though the whole attention of media now focused on the first man on the Moon.'

the resolution going beyond a random guess is not possible when only features from the four above-mentioned groups are applied. It is nevertheless true that linking *Armstrong* and *the first man on the Moon* could be easy for most human annotators — and even some search engine-based systems, even using the simplest full-text search mode.

Sometimes the situation gets complicated by the nature of the domain; the phrases *Adam Mickiewicz*, *the husband of Celina Szymanowska*, *the poet*, *the lecturer in College of France* can be clustered together only with some (deeper) knowledge about the life of Adam Mickiewicz, a Polish 19th century poet. Without referencing the history of Polish literature both a person and a computer system would experience difficulties to resolve coreference between those phrases. However, the border between common and specific knowledge is vague, especially in the face of availability of such resources as Wikipedia, offering ready-to-extract information on even less-generally known topics.

We deliberately skip one more (rare) case which should be noted for completeness: understanding of certain concepts in expert knowledge can be different from 'the common knowledge', which may hinder coreference resolution. For example, in the scientific sense tomato is the fruit (mature ovary) of the tomato plant, but in common interpretation (e.g. in cooking) tomato is a vegetable. We leave these difficult cases aside as they go beyond the scope of this paper.

3 Concept of the Pragmatic Nominal Knowledge Base

Since the nature of coreference resolution problem is conceptual: establishing and decoding coreference is about sharing the same knowledge of discourse entities between the speaker (conveying some message in the text, being the primary communication channel) and the recipient (decoding method), we could make an attempt at establishing a common, reusable, updateable platform of understanding of the facts expressed in the text being analysed. Within the scope of the resolution task, limited to nominal groups, such platform could be conceived of as a pragmatic knowledge base composed of 'seed' nominal facts and their interpretations.

This type of information goes far beyond semantic relations present e.g. in the Wordnet, with its Polish version unable to maintain definitions such as *pediatria* ('pediatrics')

— *nauka o chorobach dziecięcych* ('branch of medicine that deals with child's diseases'). Similarly, this information cannot be inferred from investigating syntactic heads of phrases since *człowiek* ('man') carries much different information capacity than the whole phrase *pierwszy człowiek na Księżycu* 'the first man on the Moon'.

The content of such base would cover established facts (such as, again, linking Neil Armstrong with his well-known attribute of being "the first man on the Moon") and typical periphrastical realisations of frequent nominal phrases, including named entities (e.g. linking Napoleon Bonaparte with his nickname, "The Little Corporal").

3.1 Data Extraction Sources

To boost development of the knowledge base, we plan to reuse existing sources of structured and unstructured data which now has been used for years in construction of semantic lexicons [6], information extraction [7] or Web question answering [8].

Structured sources should be represented by existing data- and knowledge repositories such as traditional dictionaries. For Polish two adequate resources of these type are: The Dictionary of Periphrastic Constructions [9] and The Great Dictionary of Polish — WSJP [10], both prepared by the scientific community. On the other hand, there is a growing number of crowd-sourced dictionaries and definition bases, in most cases intended to be used for Internet games and crosswords (`http://sjp.pl`, `http://krzyzowki.info`). Processing data from these groups would consist in automatic filtering of nominal definitions and passing them to manual verification.

Digital ontologies (explicit specifications of conceptualization) could also be used as source of periphrases, most likely with typical nominal instantiations of knowledge items generated in a human-readable form (to be later matched with textual content), but we deliberately omit this method, on one hand because of mostly derivative nature of such resources and on the other — due to their artificial character, abstracted from realistic use of language. To illustrate this problem, let's analyse the relation between the phrase *gród Kraka* ('Krak's (fortified) town') and its synonym, city name *Kraków* ('Cracow'). The former one is frequently used in texts about Cracow to maintain cohesion but we could hardly ever find it when looking in available structured sources. Moreover, it will never be automatically generated from any ontology because of its collocational character and atypical component *gród*, rarely used in a contemporary texts when referring to a town.

Capturing phenomena of this type can only be achieved by processing unstructured sources representing the bottom-up approach to language and likely to enrich dictionary data with real-life examples. In the long run we plan to process both balanced corpora (such as NKJP [11] or KPWr [12], providing standard representation of a language), available content sources (such as Gutenberg project) and sources of dynamic language — electronic media archives such as *Korpus Rzeczpospolitej* [13] or current parliamentary transcripts from the Polish Sejm Corpus [14].

4 The Experiment

Our hypothesis was that pragmatic data available in online data sources could improve coreference resolution in Polish by providing associations unavailable to obtain with

currently used methods (surface, syntactic, semantic or discourse-based). To verify that, we have compared manual annotation of nominal mentions in the corpus of general nominal coreference — Polish Coreference Corpus, PCC [15] with their automatic annotation created with Ruler [2] to extract coreferential links identified by human annotators, but missed by the computer resolver and having the property of semantic unrelatedness. Absence of such link gives sufficient indication that the current resolution methods could not create the association, but there exists some additional level of understanding of the text which makes it obvious for the human annotators.

Out of 1220 nominal clusters (with only nominal mentions) 73 mention pairs have been manually selected for further processing. They constituted all data for which coreference resolution was unfeasible with the above-mentioned means. Mentions which are currently not clustered, but could get resolved with additional semantic effort, were removed from the data set. Two examples of such semantic-intensive data are *czternaście tysięcy złotych* ('fourteen thousand Polish zlotys') — *pierwsza tak duża dotacja* ('the first so huge subsidy'), when cluster could have been created by comparing wordnet-based semantic classes, and *marszałek* ('marshal') — *Marek Nawara, marszałek małopolski* ('Marek Nawara, the marshal of Małopolska')[1], when appositional components could have been inspected to create the link.

4.1 Data Classification

It occurred that the contents of the set follows, to a great extent, the common classifications of named entities, such as the one used for the National Corpus of Polish (see e.g. [16]). Among the 73 mention pairs included in the set, four out of five following subclasses are named entity-related and follow the NKJP classification:

- 29 personal names linked with person role, function, occupation etc. (e.g. *Jan Paweł II* ('John Paul II') — *polski papież* ('the Polish pope'), *Rafał Blechacz — pianista ogromnie utalentowany i skromny* ('a pianist tremendously talented and modest')
- 18 names of organisations — companies, sports clubs, political parties, music bands etc. (e.g. *Ich Troje — zespół Michała Wiśniewskiego* ('Michał Wiśniewski's band'), *Wizzair — tania linia lotnicza* ('low-cost airline'))
- 14 geographical/geo-political names — here: only names of countries and cities (e.g. *Irak* ('Iraq') — *kraj* ('country'), *Aleksandrów Łódzki — miasto* ('city'))
- 6 'human creation' names — movie, book and newspaper titles (e.g. *Star Trek — dzieło filmowe* ('cinematographic work'), *Wahadło Foucaulta* ('Foucault's Pendulum') — *książka* ('a book'))
- 6 descriptive definitions, e.g. *kot* ('cat') — *udomowiony ssak* ('domesticated mammal')), *lekarze i pielęgniarki* ('doctors and nurses') — *personel szpitalny* ('hospital staff')).

Such statistics imply that the seed concepts should be closely related to named entities. It results in the first place from absence or underrepresentation of such concepts in the Polish WordNet — quoting city examples, plWordNet contains 339 sample instances of the artificial synset *miasto Polskie* ('a Polish city') which corresponds to 1/3 of

[1] In PCC appositions are treated as components of the main phrase.

the total number of all cities in Poland. Nevertheless, the structure and contents of any wordnet cannot be subordinated to ideology of representing the whole world knowledge – c.f. the Princeton WordNet, similarly far from representing company names or movie titles.

4.2 Knowledge Extraction Attempt

Each of the mention pairs have been manually tested against one of the knowledge bases mentioned in Section 3.1 to provide a proof of concept that extracted data used as 'pragmatic features' would, to a large extent, help in proper clustering of mentions in the coreference resolution process.

2 sources have been selected as main supplies of pragmatic data: Polish Wikipedia and online crossword definition service `http://krzyzowki.info`. This decision was based on the assumption that Wikipedia is a reliably enough source of information about named entities while crossword services should provide sufficient support for definitions. Table 1 provides statistics of data sources used for resolving mention pair dependency, showing number of entity pairs which could be resolved using only Wikipedia, only the crossword definitions, with both methods, some other algorithmically available method or which could not be resolved by any pragmatic means.

Table 1. Sources of pragmatic information

	Wikipedia	**krzyzowki.info**	**both**	**other**	**none**
personal names	14		14	1	
organisations	9		8		1
geo names		1	13		
creation names	1		5		
definitions	4		1	1	

The first important finding is that all but one problematic assignments could be properly resolved; the missing one resulted from manual annotation error (wrong association between a soccer club name and a mountain name: *Klimczok*). Another striking fact is that for most mention pairs (all but three) the resolution process could be completed by using only Wikipedia. The only definition-based case was the association between the country name *Niemcy* ('Germany') and its property: *zachodni sąsiad Polski* ('the western neighbour of Poland'), possible to get resolved using the textual head-match with the phrase present in the definition base.

'Other' resolution source indicates that both of the main sources were not sufficient to resolve the link, but another available online source could be used; the examples here are diminutive and augmentative form of the name *Małgorzata* ('Margaret'): *Gosia* and *Gocha* and a common name for medical staff: *lekarze i pielęgniarki* ('doctors and nurses') — *personel szpitalny* ('hospital staff').

4.3 Data Abstraction

When collected, separate set of algorithms can be used to abstract nominal facts from nominal phrases which we believe to boost coreference resolution recall while

maintaining storage efficiency. Apart from typical collocations which should only be processed in a controlled manner, two abstraction components are now envisaged: a syntactic one and semantic one.

The former would convert between different syntax models of a phrase maintaining its meaning, e.g. relative to participial phrases: *osoba, która podpowiada aktorom* ('a person who feeds lines to actors') — *osoba podpowiadająca aktorom* ('a person feeding lines to actors'). The semantic component would use wordnet relations such as synonymy or hyponymy to neutralise lexical meanings of phrase components: *osoba, która podpowiada wykonawcom* ('a person who feeds lines to performers').

Evaluation of both components would be a starting point for further investigation of several independent research problems e.g.:

- how alternation of verbal constructs influences usage of phraseology (*przejąć* ('to take over') — *dokonać przejęcia* ('to make a takeover'), *człowiek, który przepłynął Atlantyk* ('a man who sailed across the Atlantic') — *człowiek, który przebył Atlantyk* ('a man who travelled across the Atlantic'))
- how far can attributes modify nominal syntax constructs (*mała niebieska pigułka* ('little blue pill') — *niebieska pigułka* ('blue pill')
- which factors influence syntactic stability of collocations (cf. *Kraj Wschodzącego Słońca* ('Land of the Rising Sun')).

5 Conclusions and Further Work

The experiments confirmed our original hypothesis that currently available data sources can provide pragmatic knowledge and in this way improve coreference resolution in Polish when currently used algorithms fail. Apart from coreference resolution, the completed version of the database will also find its other linguistic applications such as pragmatic analysis of text for smoothing the result of automatic text summarization, machine translation or readability improvements. Development of the knowledge base would seriously enrich capabilities of independent IT systems performing text analysis, especially as current version of such systems are insensitive to pragmatic facts vital for correct interpretation of the text while such information is freely available to all search engines, even in the simplest full-text search mode (cf. search results for „Nazi Propaganda Minister"). The knowledge base could also be made available independently, in a WolframAlpha-like interface offering search and visualisation.

Considering incremental and volatile character of knowledge, expressed by constant update of underlying resources by Internet users, extraction algorithms could be linked with data sources in a way triggering updates of the knowledge base contents when source data (e.g. Wikipedia article) gets updated.

The data pool could be extended with more linked data sets and tools traditionally used for ontological modelling, with possibility of using ontological relations to improve data abstraction (e.g. when *Księżyc* ('the Moon') is linked in ontology to *Srebrny Glob* ('the Silver Globe'), it could be used in abstraction of phrases like *pierwszy człowiek na Księżycu* ('the first man on the Moon')). Interfacing with WolframAlpha or Google Knowledge Graph will be also investigated. Last but not least, foreign-language resources could be examined to import translated nominal representation of knowledge bits to the base.

References

1. Ogrodniczuk, M., Zawisławska, M., Głowińska, K., Savary, A.: Coreference Annotation Schema for an Inflectional Language. In: Gelbukh, A. (ed.) CICLing 2013, Part I. LNCS, vol. 7816, pp. 394–407. Springer, Heidelberg (2013)
2. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for Polish. In: Vetulani, Z. (ed.) Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, pp. 167–171 (2011)
3. Kopeć, M., Ogrodniczuk, M.: Creating a Coreference Resolution System for Polish. In: [17] 192–195
4. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocawskiej (2009), `http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`
5. Vetulani, Z., Kubis, M., Obrębski, T.: PolNet — Polish WordNet: Data and Tools. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) LREC. European Language Resources Association (2010)
6. Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, vol. 10, pp. 214–221. Association for Computational Linguistics, Stroudsburg (2002)
7. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Proceedings of the Fifth ACM Conference on Digital Libraries, DL 2000, pp. 85–94. ACM, New York (2000)
8. Dumais, S., Banko, M., Brill, E., Lin, J., Ng, A.: Web question answering: is more always better? In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2002, pp. 291–298. ACM, New York (2002)
9. Bańko, M.: Słownik peryfraz czyli wyrażeń omownych. PWN Scientific Publishers, Warszawa (2003)
10. Żmigrodzki, P.: O projekcie Wielkiego słownika języka polskiego. Język Polski 5(LXXXVII), 265–267 (2007)
11. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.): Narodowy Korpus Języka Polskiego (Eng.: National Corpus of Polish). Wydawnictwo Naukowe PWN, Warsaw (2012)
12. Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., Wardyński, A.: KPWr: Towards a Free Corpus of Polish. In: [17], pp. 3218–3222
13. Presspublica: Korpus Rzeczpospolitej, `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`
14. Ogrodniczuk, M.: The Polish Sejm Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey. European Language Resources Association (ELRA) (2012)
15. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In: Sun, M., Zhang, M., Lin, D., Wang, H. (eds.) CCL and NLP-NABD 2013. LNCS, vol. 8202, pp. 97–108. Springer, Heidelberg (2013)

16. Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A., Lenart, M.: Annotation tools for syntax and named entities in the National Corpus of Polish. International Journal of Data Mining, Modelling and Management 5(2), 103–122 (2013)
17. Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.): Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey. ELRA (2012)

Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents

Vishal Gupta

Computer Science & Engineering,
University Institute of Engineering & Technology,
Panjab University Chandigarh, India
vishal@pu.ac.in

Abstract. This paper concentrates on hybrid algorithm for multilingual summarization of Hindi and Punjabi documents. It combines the features of Hindi summarizer as suggested by CDAC Noida and Punjabi summarizer as suggested by Gupta and Lehal in 2012. In addition to this, it also suggests some new features for summarizing Hindi and Punjabi multilingual text. It is first time that this multilingual text summarizer has been proposed which supports both Hindi and Punjabi text. Nine features used in this algorithm for summarizing multilingual Hindi and Punjabi text are: 1) Key phrase extraction 2) Font feature 3) Nouns and Verbs Extraction 4) Position feature 5) Cue-phrase feature 6) Negative keywords extraction 7) Named Entities extraction 8) Relative length feature 9) extraction of number data. For each sentence, scores of each feature is calculated and then machine learning based mathematical regression is applied for identifying weights of these nine features. Sentence final-scores are calculated from feature weight equations. Top scored sentences in proper order (in same order as in input) are selected for final summary. Default summary is made at 30% compression ratio. This algorithm performs well at 30% compression ratio for both intrinsic and extrinsic measures of summary evaluation. This algorithm has been thoroughly tested on 30 Hindi-Punjabi documents and reports F-Score equal to 92.56% which is reasonably good.

Keywords: Hybrid Multilingual Summarizer, Multilingual Hindi Punjabi Summarizer, Hindi Extractive Summarization, Punjabi Extractive Summarizer.

1 Introduction

Automatic text summarization [1] involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the meaning of the text. The brief summary produced by summarization system allows readers to quickly and easily understand the content of original documents without having to read each individual document. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. Text summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method [2] deals with

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 717–727, 2013.
© Springer International Publishing Switzerland 2013

selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original document.

Microsoft Word's AutoSummarize [2] function is a simple example of text summarization system for English. In some of summarization systems, users can specify percentage of total source text in final summary. Worthiness of lengthy documents can quickly and easily be judged using text summarization. A good summarization system should reflect the diverse topics of the document while keeping redundancy to minimum. Automatic text summarization is one of the widely used applications in the field of natural language processing (NLP) [3]. Various automatic text summarization systems are commercially or non-commercially available for most of the commonly used natural languages. Most of these text summarization systems are for English and other foreign languages. Moreover, for commercial products the technical documentation is often minimal or even absent. When it comes to Indian languages, automatic text summarization systems are still lacking.

This paper concentrates on hybrid algorithm for multilingual extractive summarization of Hindi and Punjabi text. It combines the features of Hindi Summarizer as suggested by CDAC Noida [10] and Punjabi summarizer as suggested by Gupta et al. (2012) [4]. In addition to this, it also suggests some new features for summarizing Hindi and Punjabi multilingual text. It is first time that this hybrid algorithm for multilingual text summarization has been proposed which supports both Hindi and Punjabi text. Various features used in this algorithm for summarizing multilingual Hindi and Punjabi text are: 1) Key phrase extraction 2) Font feature 3) Nouns and Verbs Extraction 4) Position feature 5) Cue-phrase feature 6) Negative keywords extraction 7) Named Entities extraction 8) Relative length feature 9) extraction of number data. For each sentence, scores of each feature is calculated and then machine learning based mathematical regression is applied for identifying weights of these nine features. Sentence final-scores are calculated from feature weight equations. Top scored sentences in proper order (in same order as in input) are selected for final summary. Default summary is made at 30% compression ratio.

2 Features Used in Hybrid Summarization Algorithm

There are nine features used in hybrid algorithm for multilingual summarization of Hindi and Punjabi text: 1) Key phrase extraction 2) Font feature 3) Nouns and Verbs Extraction 4) Position feature 5) Cue-phrase feature 6) Negative keywords extraction 7) Named Entities extraction 8) Relative length feature 9) extraction of number data. Before calculating these features first of all boundary is identified for Hindi-Punjabi sentences and words [8]. Then Hindi-Punjabi Stop words are removed from sentences. A list of Hindi-Punjabi stop words has been manually developed. Stop words are un-important words with very high frequency. Then we have applied Punjabi stemmer for nouns and proper nouns as developed by Gupta et al. (2011) [11]

and for Hindi, a lightweight stemmer proposed by Ramanathan and Rao (2003) [12] has been applied.

This lightweight stemmer proposed for Hindi is based on the grammar for Hindi language in which a list of total 65 suffixes is generated manually. Terms are conflated by stripping off word endings from a suffix list on a `longest match' basis. Noun, adjective and verb infections have been discussed and based on that 65 unique suffixes are collected. The major advantage of this approach is as it is computationally inexpensive. In Punjabi stemmer an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun morph and proper names list and various stemming rules for nouns and proper names are generated. Various features used for summarizing multilingual Hindi-Punjabi text are given below:

2.1 Extraction of Key Phrases from Hindi and Punjabi Sentences

Key phrases are the thematic words which can represent the theme of whole text. Those sentences which possess these thematic words are important for summary. Various Hindi-Punjabi noun words which are in bold, italics or underlined font or having high scores for TF (Term-Frequency) [7] or which appear in title lines (i.e. title keywords) are important and are candidates of key phrases from this phase. These words are treated as Hindi-Punjabi key phrases. Hindi word-net [13] has been used for finding Hindi nouns. For extracting Punjabi nouns Punjabi noun-morph is used which is having 37297 Punjabi nouns. Term frequency is number of times a Hindi-Punjabi noun appears in input text. If TF-Score of a noun is more then it means that it is more important word. Normally, words written in bold, italics or underlined fonts or having more font size are more important than the other words. We identify the Hindi-Punjabi nouns written in bold, italics or underlined fonts. Along with this we also identify the Hindi-Punjabi noun words having more font size than the normal size used for regular text. Title lines are the headlines of documents. Those Hindi-Punjabi sentences containing title keywords [6] are important. Title keywords are obtained after removing Hindi-Punjabi stop words from title lines. Hindi-Punjabi key phrases are extracted from this phase and those sentences in Hindi-Punjabi which contain these key phrases are important. The score of this feature is calculated by dividing number of unique key phrases in a sentence with total number of key phrases.

2.2 Font Feature for Hindi-Punjabi Sentences

Those Hindi-Punjabi sentences, which are in bold, italics or underlined font or having more font size than the rest of text are more important and are candidates for summary sentences [5]. Usually headlines of newspapers, title lines of stories, title line of articles and other important sentences appear in bold, italics or underlined font or with more font size. These title lines and headlines are sufficient to describe the sense of whole document. This feature [6] is the most important feature in case of Hindi-Punjabi Summarizer. Those sentences which come as output of this phase are candidates for summary sentences and their font-feature flag is set to true.

2.3 2.3 Extraction of Hindi-Punjabi Nouns and Verbs

Those Punjabi sentences containing Hindi-Punjabi nouns and verbs [1] are important. In case of Hindi, input words are checked in Hindi Wordnet [13] for the possibility of Hindi nouns and verbs. In case of Punjabi, words are checked in Punjabi noun morph for possibility of nouns and Punjabi dictionary for the possibility of Punjabi verbs. Moreover common Hindi-English nouns and common Punjabi-English nouns are also searched in this phase from their respective dictionaries of common Hindi-English nouns and common Punjabi-English nouns. Examples of common Hindi-English and Punjabi-English words are 'टैकनॉलॉजी' ਟੈਕਨਾਲੋਜੀ and 'मोबाइल' ਮੋਬਾਈਲ. The score of this feature is determined by dividing number of Hindi-Punjabi nouns in a sentence with length of that sentence. The value of this feature for a sentence will be from 0 to 1. If value for this feature for a particular sentence is closer to 1, then it means, that sentence is more important.

2.4 Position Feature for Hindi-Punjabi Sentences

Usually first sentence of first paragraph and last sentence of last paragraph are more important [6] as they can contain lot of information about the topic than rest of text. Set the flag for Position feature true for those Hindi-Punjabi sentences which belong to first or last sentence of first or last paragraphs respectively.

2.5 Extraction of Hindi-Punjabi Cue Phrases

Cue Phrases [1] are certain phrases like In Conclusion, Summary and Finally etc. Those sentences which are beginning with cue phrases or which contain these cue phrases are generally more important than others. Initially a list of Hindi-Punjabi Cue phrases has been made and then those sentences containing these cue phrases are given more importance. For example some of Hindi-Punjabi cue phrases are 'अंत में' ਅੰਤ ਵਿੱਚ "in the end" and 'संक्षिप्त में' ਸੰਖੇਪ ਵਿੱਚ "in brief" etc. Set cue phrase flag to true for those Hindi-Punjabi sentences which contain these Hindi-Punjabi cue phrases.

2.6 Extraction of Non Important Information from Hindi-Punjabi Sentences

Some words are indicators of non-essential information [1] and are called negative keywords. These words are speech markers such as Hindi-Punjabi words 'क्योंकी' ਕਿਉਂਕੀ "because" '', ਹੋਰ ਵੀ 'और भी' "furthermore", and 'के अतिरिक्त' ਦੇ ਇਲਾਵਾ "additionally", and 'आम तौर पर' ਆਮ ਤੌਰ ਤੇ "typically" occur in the beginning of a sentence. This is a binary feature, taking on the value 0 or 1. The flag for this feature is set if a sentence contains at least one of these discourse markers, and "false" otherwise.

2.7 Extraction of Hindi-Punjabi Named Entities

Named Entities are very important. Those sentences which contain these named entities are important for summary. For Hindi, we have applied CRF based named entity recognition system developed by Sharma et al. (2011) [14]. For Punjabi we have applied rule based Punjabi named entity recognition system [15] developed by Gupta et al. (2011). In CRF approach for Hindi named entity recognition, 12 NE tags are used. Six features: context word feature, word prefix, word suffix, POS information, NE feature and various Gazetteer lists have been used. The contextual window of size seven, prefix and suffix length and NE information of the previous word, current word and different features have been used. The Precision, Recall and F-Score for this system are 72.78%, 65.82% and 70.45% respectively. Punjabi rule based named entity recognition system uses various gazetteer lists like prefix list, suffix list, middle name list, last name list and proper name lists for checking whether the given word is proper name or not. After doing analyses of Punjabi corpus of Ajit Punjabi newspaper, various gazetteer lists have been developed. For Punjabi NER, Precision=89.32%, Recall=83.4% and F-score=86.25%. The score for this feature is calculated by dividing number of Hindi-Punjabi named entities in a sentence with length of that sentence. The value of this feature for a sentence is calculated by taking ratio of number of Hindi-Punjabi named entities in a sentence to the length of that sentence. For example in the following sentence there are two named entities. ज़िला बरनाला के डिप्टी कमिशनर स: अरशदीप सिंह ने अहम मीटिंग की। ਜ਼ਿਲ੍ਹਾ ਬਰਨਾਲਾ ਦੇ ਡਿਪਟੀ ਕਮਿਸ਼ਨਰ ਸ: ਅਰਸ਼ਦੀਪ ਸਿੰਘ ਨੇ ਅਹਿਮ ਮੀਟਿੰਗ ਕੀਤੀ।.

The two Named entities are: 'बरनाला' ਬਰਨਾਲਾ and 'स: अरशदीप सिंह' ਸ: ਅਰਸ਼ਦੀਪ ਸਿੰਘ.

2.8 Relative Length Feature for Hindi-Punjabi Sentences

Hindi and Punjabi are very closely related languages and in both the languages short sentences are avoided [5] for including in final summary as often they contain less information. On the other hand lengthy Hindi-Punjabi sentences might contain lot of information. This feature is calculated by dividing number of Hindi-Punjabi words in a sentence with word count of largest sentence. Its value for this feature lies between 0 to 1.

Hindi-Punjabi-Sentence-Relative-Length Score= number of Hindi-Punjabi words in a sentence / word count of largest sentence.

2.9 Numeric Data Identification

The sentence that contains numeric data [5] is important and it is most probably included in the document summary. Numeric digits, Devanagari numerals for Hindi

(०,१,२,३,४,५,६, ७,८,९), Gurmukhi numerals for Punjabi (੦,੧,੨,੩,੪,੫,੬,੭,੮,੯.) and Roman numerals (i, ii,iii, iv, v, vi, vii, viii, ix, x, xi and xii etc.) are considered as numeric data. The score for this feature is calculated as the ratio of the number of numeric data items in a sentence by the sentence length.

3 Hybrid Algorithm for Hindi-Punjabi Text Summarization

Finally actual scores of sentences are determined from sentence-feature-weight equation: $w_1f_1+w_2f_2+ w_3f_3+\ldots\ldots\ldots\ldots\ldots w_9f_9$ Where $f_1, f_2, f_3\ldots\ldots\ldots\ldots f_9$ are nine features of Hindi-Punjabi sentences calculated in the different sub phases and $w_1, w_2, w_3\ldots\ldots\ldots\ldots w_9$ are the corresponding feature weights of sentences. Mathematical regression [5] has been used as model to estimate the text features weights for Hindi-Punjabi text summarization.

A relation between inputs and outputs is established. In matrix notation we can represent regression as follow:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ . \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & .... & X_{09} \\ . & . & .... & . \\ . & . & .... & . \\ X_{m1} & X_{m2} & .... & X_{m09} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ . \\ w_m \end{bmatrix}$$

Where

$[Y]$ is fuzzy output vector having values between 0 to 1 based on importance of sentences given manually for 30 Hindi-Punjabi documents.

$[X]$ is the input matrix (feature parameters) for nine features having values between 0 to 1.

$[w]$ is linear statistical model of system (the weights w_1, w_2…….w_9 in the equation)

m is total number of sentences in the training corpus.

Weight w of a particular feature k (k=1 to 9) with input matrix x and fuzzy output matrix y can be calculated as follows:

$$w = \frac{\sum_{i=01 \text{ to } m1} (x_i - \text{mean}(x))(y_i - \text{mean}(y))}{\sum_{i=01 \text{ to } m1} (x_i - \text{mean}(x))^2}$$

From the above equation, weights of each of nine features of Hindi-Punjabi text summarization have been calculated. Results of weight learning for nine features are shown in Table 1.

Table 1. Weight learning results using regression

Features	Learned weights
Key phrase extraction	+2.86
Font feature	+4.47
Nouns and Verbs extraction	+1.95
Position feature	+3.77
Cue-phrase feature	+0.87
Negative keywords extraction	-1.25
Named Entities extraction	+2.34
Relative length feature	+0.72
Extraction of number data.	+0.90

From results of weight learning in Table 1, we can conclude that two most important features of hybrid algorithm for multilingual summarization of Hindi and Punjabi text are font feature and position feature. Top scored Hindi-Punjabi sentences in proper order (in order of sentences in input) are selected for final summary at 30% compression ratio which is default compression ratio.

Hybrid Algorithm for multilingual Hindi-Punjabi Text Summarization:-

This Hybrid algorithm proceeds by segmenting the source Hindi-Punjabi text into sentences and words.

Step 0: Set the scores of each sentence as 0.

Step 1: Delete the duplicate Hindi-Punjabi sentences from input text by searching the current sentence in the sentence list which is initially empty. For each sentence check the following condition: If current sentence is found in sentence list then Current sentence is set to null being the duplicate sentence. Else Current sentence is added to the sentence list being the unique sentence.

Step 2: Delete all the occurrences of Hindi-Punjabi stop word from input text.

Step 3: Apply Hindi-Punjabi stemmer for converting Hindi-Punjabi words into their root words.

Step 4: Calculate the key phrase feature score for all the Hindi-Punjabi sentences.

Step 5: Calculate the font feature score for all Hindi-Punjabi sentences.

Step 6: Calculate the Hindi-Punjabi nouns and verbs extraction feature score for all the sentences.

Step 7: Calculate Position feature score for all the Hindi-Punjabi sentences.

Step 8: Calculate Cue-phrase feature score for all the Hindi-Punjabi sentences.

Step 9: Calculate the Score of non essential information feature i.e. negative keywords feature score for all the Hindi-Punjabi sentences.

Step 10: Calculate the score of named entity extraction feature for all Hindi-Punjabi sentences.

Step 11: Calculate the relative length feature score for all the Hindi-Punjabi sentences.

Step 12: Calculate Number feature score for all Hindi-Punjabi sentences.

Step 13: Calculate the weight-age of each feature by applying regression using sentence-feature-weight-equation.

Step 14: Calculate final-scores of all the sentences by applying sentence-feature-weight-equation.

Step 15: Select the top scored sentences at 30% compression ratio.

Step 16: Final multilingual Hindi-Punjabi summary is formed by arranging top scored sentences in ascending order of their position in input text at 30% compression ratio.

Algorithm Input:

प्यारा सा घर

हर मनुष्य की इच्छा होती है कि उसका एक प्यारा सा घर हो। कई लोग उम्र भर की कमाई मकान बनाने में लगा देते हैं। चाव के साथ सुंदर इमारत तैयार करवाई जाती है। उस में कीमती पर्दे, सोफे और ओर सजावट का समान लगाया जाता है। फ़रिज, टैलिविज़न और एयर-कंडीशनर आदि किसी किस्म की कमी नहीं रहने दी जाती। ऐसे मकान बहुत सुंदर लगते हैं और दूर से ही मन को मौंह लेते हैं। ਪਰ ਜੇ ਇਨਾਂ ਵਿਚ ਪਿਆਰ ਦੀ ਘਾਟ ਹੋਵੇ ਤਾਂ ਮਕਾਨ ਮਕਾਨ ਹੀ ਰਹਿੰਦੇ ਹਨ। ਕਦੇ ਘਰ ਨਹੀਂ ਬਣ ਸਕਦੇ ਕਿਉਂਕਿ ਘਰ ਇੱਟਾਂ, ਪੱਥਰ ਅਤੇ ਸੰਗਮਰਮਰ ਦੀ ਇਮਾਰਤ ਨੂੰ ਨਹੀਂ ਕਹਿੰਦੇ। ਘਰ ਅਤੇ ਮਕਾਨ ਵਿਚ ਬਹੁਤ ਅੰਤਰ ਹੁੰਦਾ ਹੈ। ਘਰ ਬਣਦਾ ਹੈ ਕਿਸੇ ਇਮਾਰਤ ਵਿਚ ਰਹਿਣ ਵਾਲੇ ਬੰਦਿਆਂ ਦੇ ਪਰਸਪਰ ਪਿਆਰ ਨਾਲ।

Algorithm Output at 30% Compression Ratio:

प्यारा सा घर

हर मनुष्य की इच्छा होती है कि उसका एक प्यारा सा घर हो। ਘਰ ਬਣਦਾ ਹੈ ਕਿਸੇ ਇਮਾਰਤ ਵਿਚ ਰਹਿਣ ਵਾਲੇ ਬੰਦਿਆਂ ਦੇ ਪਰਸਪਰ ਪਿਆਰ ਨਾਲ।

Text in blue line indicates title line. Text in black color indicates Hindi text. Text in Red color indicates Punjabi text.

4 Results and Discussions

The hybrid algorithm for Hindi-Punjabi text summarizer has been tested over 30 Hindi-Punjabi documents including general articles and stories. Data set contains 2314 Hindi-Punjabi sentences and 20352 Hindi-Punjabi words. These 30 Hindi-Punjabi documents were collected from Popular Punjabi websites www.likhari.org & Punjabi news paper http://www.ajitjalandhar.com/ and Hindi news papers http://www.jagran.com & http://www.amarujala.com/. We have applied four intrinsic measures [9] of summary evaluation 1) F-Score 2) Cosine Similarity 3) Jaccard Coefficient and 4) Euclidean distance and one extrinsic measure of summary evaluation: Question Answering Task.

$$\text{F-Score} = \frac{(2 * \text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

$$\text{Recall} = \frac{\text{Number of correct sentences retrieved by our system}}{\text{Total number of sentences retrieved by human expert}}$$

$$\text{Precision} = \frac{\text{Number of correct sentences retrieved by our system}}{\text{Total number of sentences retrieved by our system.}}$$

Given two documents and vectors A and B are the term frequency vectors of these documents over the term set T= $\{t_1,\ldots\ldots, t_m\}$ Cosine similarity between two vectors is calculated as follows: COSINE_SIMILARITY (A, B) = Cos(Θ)= $(A \cdot B) / (|A| |B|)$

$$= \sum A_i \times B_i / \sqrt{\sum (A_i)^2} \times \sqrt{\sum (B_i)^2} \text{ where i= 1 to n}$$

Jaccard Coefficient = SIM (A, B) = $(A \cdot B)/ (|A|^2 + |B|^2 - A \cdot B)$

$$= (A \cdot B) / (\sqrt{\sum (A_i)^2} \times \sqrt{\sum (A_i)^2} + \sqrt{\sum (B_i)^2} \times \sqrt{\sum (B_i)^2} - A \cdot B)$$

Where i= 1 to n and each dimension represents a term with its frequency in the document. Jaccard Coefficient is a similarity measure and ranges from 0 to 1.

The Euclidean distance of the two documents is defined as:

Euclidean distance$(X_{ik}, X_{jk}) = (\sum (X_{ik} - X_{jk})^2)^{1/2}$ for k=1 to n key terms.

Firstly we have produced gold summaries (reference summaries) of these 30 Hindi-Punjabi documents. For making the gold summaries, two human experts (having good knowledge of Hindi and Punjabi) have been assigned the task of producing the manual summaries separately of these 30 documents at 30% compression ratio. Finally gold summaries (reference summaries) are produced by including mostly common sentences of two manual summaries produced by two human experts at 30% compression ratio. The results of intrinsic summary evaluation are shown in Table 2.

Table 2. Intrinsic summary evaluation for Hindi-Punjabi text summarizer

Compression ratio (In %)	Intrinsic summary evaluation			
	Avg. F-score	Avg. Cosine similarity	Avg. Jaccard coeff.	Avg. Euclidean distance
30%	92.56	0.910	0.902	0.46

As can be seen from Table 2, for intrinsic measures of summary evaluation, Hindi-Punjabi Text summarizer is reasonably performing well at 30% compression ratio.

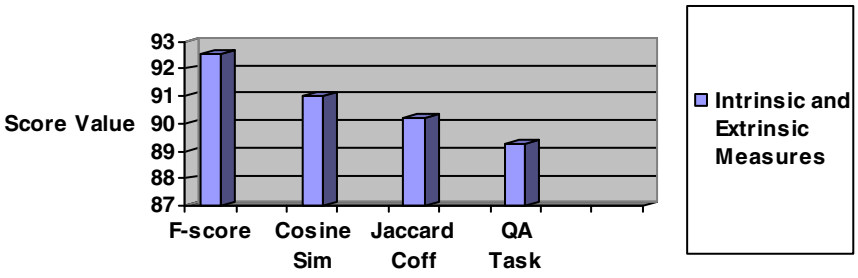


Fig. 1. Intrinsic and extrinsic measures for multilingual Hindi-Punjabi summarizer

Figure 1 shows the intrinsic and extrinsic measures of summary evaluation for hybrid algorithm of multilingual Hindi-Punjabi text summarizer. For performing the task of question answering, firstly two human experts have been given 30 Hindi-Punjabi documents and then they jointly prepared five questions for each of 30 documents. Then answers of these questions are looked into system produced

summary. For each correct answer, counter for number of correct answers is incremented by one for that document. Accuracy for performing task of question answering is calculated as follows:

Accuracy= No. of correct answers/ Total No. of questions asked

For Hindi-Punjabi text summarizer, accuracy of question answering task is 89.25% at 30% compression ratio which seems to be reasonably well.

No performance measure (intrinsic or extrinsic measure) has been reported by Hindi summarizer suggested by CDAC Noida [10]. On the other hand Punjabi summarizer suggested by Gupta et al. (2012) [4] reported accuracy of only 84.26% for performing question answering task as extrinsic measure of summary evaluation at 30% compression ratio for Punjabi stories. Hence we can conclude hybrid algorithm for multilingual summarization of Hindi and Punjabi documents performs reasonable well as compared to both Hindi and Punjabi summarizers.

5 Conclusions

The algorithm for multilingual summarizer is hybrid algorithm for summarizing Hindi and Punjabi text. It is first of its kind algorithm as no other summarizer exists in the world which supports both Hindi and Punjabi text. It combines the features of Hindi Summarizer as suggested by CDAC Noida [10] and Punjabi summarizer as suggested by Gupta et al. (2012) [4]. In addition to this, it also suggests some new features for summarizing Hindi and Punjabi multilingual text. For developing this multi lingual summarizer various basic linguistic resources for Hindi and Punjabi have been developed from scratch like Hindi-Punjabi stop words list, Hindi-Punjabi prefix names list, suffix names list, last names list, middle names list, Hindi-Punjabi stemming rules and Hindi Punjabi cue phrase list etc.

This summarizer can save the time of people for reading lengthy Hindi-Punjabi documents by quickly producing summary. For both intrinsic and extrinsic measures of summary evaluation, it performs reasonably well as compared to Hindi summarizer and Punjabi summarizer.

References

1. Kyoomarsi, F., Khosravi, H., Eslami, E., Dehkordy, P.K.: Optimizing Text Summarization Based on Fuzzy Logic. In: IEEE International Conference on Computer and Information Science, pp. 347–352. University of Shahid Kerman, UK (2008)
2. Gupta, V., Lehal, G.S.: A Survey of Text Summarization Extractive Techniques. International Journal of Emerging Technologies in Web Intelligence 2, 258–268 (2010)
3. Lin, J.: Summarization. In: Encyclopedia of Database Systems. Springer, Heidelberg (2009)
4. Gupta, V., Lehal, G.S.: Automatic Punjabi Text Extractive Summarization System. In: International Conference on Computational Linguistics, COLING 2012, pp. 191–198. IIT Bombay, India (2012)
5. Fattah, M.A., Ren, F.: Automatic Text Summarization. World Academy of Science Engineering and Technology 27, 192–195 (2008)

6. Kaikhah, K.: Automatic Text Summarization with Neural Networks. In: IEEE International Conference on Intelligent Systems, Texas, USA, pp. 40–44 (2004)
7. Neto, J.L., Santos, A.D., Kaestner, C.A.A., Freitas, A.A.: Document Clustering and Text Summarization. In: International Conference on Practical Application of Knowledge Discovery & Data Mining, London, pp. 41–55 (2000)
8. Gupta, V., Lehal, G.S.: Complete Pre processing Phase of Punjabi Language Text Summarization. In: International Conference on Computational Linguistics, COLING 2012, pp. 199–205. IIT Bombay, India (2012)
9. Hassel, M.: Evaluation of Automatic Text Summarization. Licentiate Thesis, Stockholm, Sweden, pp. 1–75 (2004)
10. http://www.cdacnoida.in/snlp/digital_library/text_summ.asp
11. Gupta, V., Lehal, G.S.: Punjabi Language Stemmer for Nouns and Proper Names. In: Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) IJCNLP 2011, Chiang Mai, Thailand, pp. 35–39 (2011)
12. Ramanathan, A., Rao, D.D.: A Lightweight Stemmer for Hindi. In: Proceedings of Workshop on Computational Linguistics for South-Asian Languages. EACL (2003)
13. http://www.cfilt.iitb.ac.in/wordnet/webhwn/
14. Sharma, R., Goyal, V.: Name Entity Recognition Systems for Hindi Using CRF Approach. In: Singh, C., Singh Lehal, G., Sengupta, J., Sharma, D.V., Goyal, V. (eds.) ICISIL 2011. CCIS, vol. 139, pp. 31–35. Springer, Heidelberg (2011)
15. Gupta, V., Lehal, G.S.: Named Entity Recognition for Punjabi Language Text Summarization. International Journal of Computer Applications 33, 28–32 (2011)

Identifying Psychological Theme Words from Emotion Annotated Interviews

Ankita Brahmachari, Priya Singh, Avdhesh Garg, and Dipankar Das

Department of Computer Science and Engineering, NIT Meghalaya, India
{ankita.brahmachari,priyasinghn,avdheshgarg16,
dipankar.dipnil2005}@gmail.com

Abstract. Recent achievements in Natural Language Processing (NLP) and Psychology invoke the challenges to identify the insight of emotions. In the present study, we have identified different psychology related theme words while analyzing emotions on the interview data of ISEAR (International Survey of Emotion Antecedents and Reactions) research group. Primarily, we have developed a Graphical User Interface (GUI) to generate visual graphs for analyzing the impact of emotions with respect to different background, behavioral and physiological variables available in the ISEAR dataset. We have discussed some of the interesting results as observed from the generated visual graphs. On the other hand, different text clusters are identified from the interview statements by selecting individual as well as different combinations of the variables. Such textual clusters are used not only for retrieving the psychological theme words but also to classify the theme words into their respective emotion classes. In order to retrieve the psychological theme words from the text clusters, we have developed a rule based baseline system considering unigram based keyword spotting technique. The system has been evaluated based on a *Top-n* ranking strategy (where *n=10*, *20* or *30* most frequent theme words). Overall, the system achieves the average F-Scores of *.42, .32, .36, .42, .35, .40* and *.40* in identifying theme words with respect to *Joy, Anger, Disgust, Fear, Guilt, Sadness* and *Shame* emotion classes, respectively.

Keywords: Theme Word, Psychology, Emotions, Symptoms, Interview.

1 Introduction

Languages, text or words are the medium by which cognitive, clinical and social psychologists attempt to understand the human beings. But, the extraction of emotions from texts is not an easy task due to its restricted access in case of objective observation or verification [1]. Moreover, the same textual content can be presented with different emotional slants [2]. Ekman [3], for instance, derived a list of six basic emotions from facial expressions which were employed as the classes in most of the affect recognition tasks [4]. The analysis of sentiments or emotions in texts has wide range of applications. It can be used in developing new methodologies for classification of evaluative expressions at word, phrase and sentence level as well as for tracking of sentiments also [10].

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 728–739, 2013.
© Springer International Publishing Switzerland 2013

In the very beginning, the computerized text analysis in psychology was proposed by Philip Stone and his colleagues [5]. One limitation of this approach was that it relied only on the manipulation and weighting of language variables that are not visible to the user. The first truly transparent text analysis method was pioneered in [6] where the everyday words that people use (e.g., the words such as *pronouns* and *articles* etc.) were explored. Over the span of a decade, the author hand-counted people's words in texts such as political speeches and medical interviews and noticed that the first-person singular pronouns (e.g., *I*, *me*, *my*) were reliably linked to people's different levels of *depression*. But, as far as our knowledge to date, no efficient computational model is available to identify the psychiatric symptoms of a person from related texts by analyzing his/her emotions with respect to time, intensity, event etc.

Thus, our present study is mainly devoted to recognize the possible emotional outcomes of the psychiatric patients based on different background, physiological and behavioral variables. The present system generates the statistics and graphs of the emotions with respect to *sex, age, country, city, religion* and various other variables related to the patients interview data. Such variables are helpful in determining how the emotional state of a person is normal or how much it is deflected. On the basis of such results, it can be easier to predict the reactions of a person from the text data provided by him or her. Moreover, such kind of system is useful for the medical practitioners for clustering the people of similar emotions and psychological symptoms and treating them accordingly.

In the context of analyzing psychological variables and their impacts on emotions, we found very less number of interesting tasks in the literature as described in [8], [9]. In contrast to their attempts either in data mining or in sentic computing, we have identified the theme words for psychological analysis of the patients and the practitioners. Moreover, we have developed a GUI based data analyzer for analyzing emotions of the patients based on general, behavioral as well as physiological variables related to human psychology.

In the present attempt, we have used a rule based approach for identifying the psychological theme words. The derived results from a long questionnaire are important for retrieving the focused texts or emotions of a patient. For our analysis, we have used the corpus from the International Survey of Emotion Antecedents and Reactions (ISEAR) dataset [7]. The survey was conducted in 1990s across 37 countries and had almost about 3000 respondents. The graphical interface generates different visual graphs for analyzing the impacts of emotions with respect to different background, behavioral and physiological variables. At first, we have discussed the interesting results as observed from the visual graphs generated with respect to the variables and secondly, identified the text clusters from the interview statements by selecting individual and different combinations of such variables. The text clusters are used not only for retrieving the theme words but also to classify the theme words into their respective emotion classes. We have developed a rule based baseline system considering unigrams to retrieve the psychological theme words. We have ranked the theme words based on the *Top-n* ranking strategy where the values of *n*, i.e. the number of most frequent words are chosen as 10, 20 and 30, respectively. Overall, the system achieves satisfactory average F-scores with respect to all emotion classes in *Top-30* ranking strategy.

The present manuscript is divided into five different Sections. The Section 2 contains a brief description of the ISEAR dataset whereas the Section 3 discusses the graphical and visual analysis of patients' emotions based on different background, physiological and behavioral variables. The rule based approach of clustering emotional statements and identification of psychological theme words are described in the Section 4. Finally, Section 5 concludes the paper by describing the future approaches to be taken into consideration.

2 Corpus Description

The ISEAR (INTERNATIONAL SURVEY ON EMOTION ANTECEDENTS AND REACTIONS) dataset has been considered in our present task to accomplish our research goals. The dataset was concerned with different types of emotional experiences that people have in everyday life. During the survey, they asked people to recall some occasions on which they have experienced one of the following emotions: *joy, fear, anger, sadness, disgust, shame* and *guilt*. For each of these emotions, the respondents were asked to think of a situation which these feelings aroused in them and for which they vividly remember both the circumstances and their reactions. The ISEAR dataset contains psychological statements of about 3~4 sentences pre-classified into the above mentioned seven categories of emotion. Different types of numerous variables were used in their questionnaire. We have tried to analyze some of the variables from the perspective of emotions in our present approach. A GUI based interface as shown in Figure 1 has been developed to analyze the visual graphs generated on the basis of different psychological variables and their various combinations. The variables, their values and meanings available in the questionnaire are mentioned in Tables 1 and Table 2, respectively.

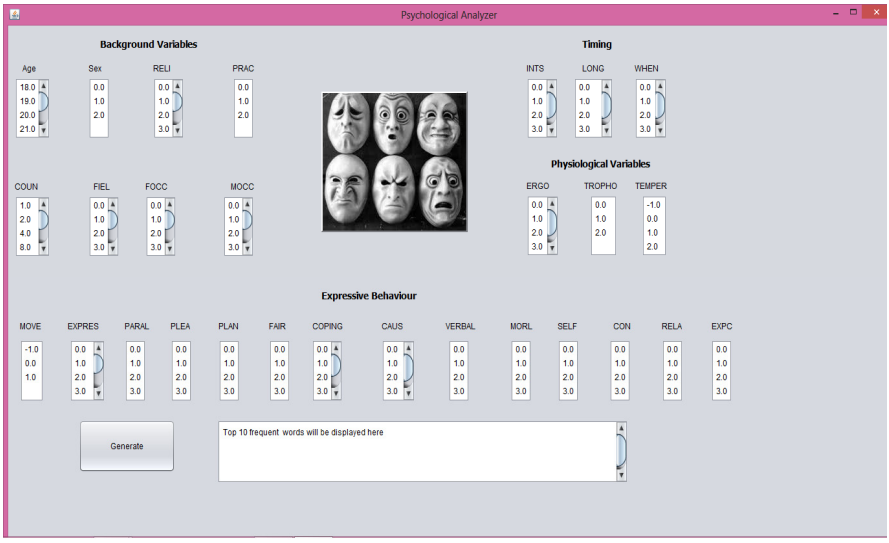


Fig. 1. Graphical User Interface (GUI) based Psychological Emotion Analyzer

Table 1. Values and corresponding meanings of different variables in the ISEAR Dataset

Background Variables	**(Value, Meaning)**
SEX	(1, MALE), (2, FEMALE)
RELI	(1, PROTESTANT), (2, CATHOLIC), (3, JEWISH), (4, HINDU), (5, BUDDHIST), (6, NATIVE), (7, OTHERS), (8, ARELIGIOUS)
PRAC	(1, PRACTISING), (2, NOT PRACTISING)
AGE	Values ranging from 18 to 35 corresponding to age
General Variables	**(Value, Meaning)**
INTS	(1, not very), (2, moderately intense), (3, Intense), (4, very intense)
WHEN	(1, days ago), (2, weeks ago), (3, months ago), (4, years ago)
LONG	(1, a few minutes), (2, an hour), (3, several hours), (4, a day or more)
Physiological Variables	**(Value, Meaning)**
ERGO (Ergotropic Arousal)	(1, change in breathing), (2, heart beating faster), (3, muscles tensing/trembling), (4, perspiring/moist hands)
TROPHO (Trophotropic Arousal)	(0, lump in throat), (1, stomach troubles) (2, crying/sobbing)
TEMPER (Felt temperature)	(-1, feeling cold/shivering), (0, no temperature symptom), (1, feeling warm/pleasant), (2, feeling hot/cheeks burning)
Expressive Variables	**(Value, Meaning)**
MOVE (Movement behavior)	(-1, withdrawing), (0, no movement), (1, moving towards people and things)
EXPRES (Nonverbal activity)	(0, no activity), (1, laughing/smiling), (2, crying/sobbing), (3, other facial expression change), (4, screaming/yelling), (5, other voice changes), (6, changes in gesturing)
PARAL (Paralinguistic activity)	(0, no activity), (1, speech-melody change), (2, speech disturbances), (3, speech tempo change)
VERBAL (verbal activity)	(0, silence), (1, short utterance), (2, one or two sentences), (3, lengthy utterance)

Table 2. Values and meanings of different variables with respect to questions

Question	**Variable**	**(Value, Meaning)**
1. Did you try to hide or to control your feelings so that nobody would know how you really felt?	CON	(0, not applicable), (1, not at all) (2, a little), (3, very much)
2. Now please think back to the situation or event that caused your emotion. Did you expect this situation to occur?	EXPC	(0, not applicable), (1, not at all), (2, a little), (3, very much)
3. How did this event affect your feelings about yourself, such as your self-esteem or your self confidence?	SELF	(0, not applicable), (1, negatively), (2, not at all), (3, Positively)
4. How did this event change your relationships with the people involved?	RELA	(0, not applicable), (1, negatively), (2, not at all), (3, Positively)

3 Graphical Analysis of Emotions Based on Psychological Variables

The statistics and graphs that were generated using the system show some interesting results in a comparative analysis of all the seven emotions. Analysis of graphs and their observations are described in the next sub-sections. The following seven color codes have been used to show seven different emotions in the graphs.

3.1 Ergotropic and Trophotropic Arousals with Respect to Emotions

We have observed that the change in breathing is slower in case of *sadness* and *guilt* whereas faster in case of *anger, joy* and *shame* (as shown in Figure 2). In case of *fear* and *anger*, the heart beat becomes faster whereas it is normal in case of *guilt*. Again, in case of *fear* and *anger*, muscles are tensed but they are relaxed in case of *joy*. Perspiring and moist hands are seen maximum in case of *fear* comparing to other emotions.

One interesting observation as shown in Figure 3 is that the *lump in throat* is high in case of *joy* and low in case of *sadness* whereas *stomach troubles* are high in case of *sadness* and low in case of *joy*. *Crying* and *sobbing* are maximum in case of *sadness* and *fear* whereas minimum in case of *joy*.

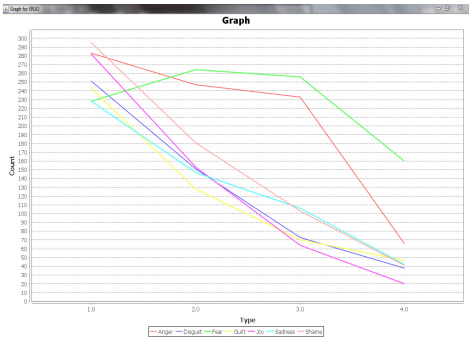


Fig. 2. Changes of Seven Emotions with respect to Different Ergotropic Arousals (in Table 1)

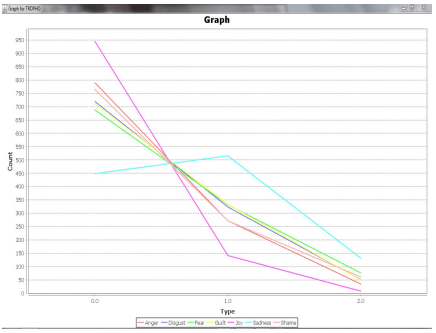


Fig. 3. Changes of Seven Emotions with respect to Different Trophotropic Arousals (in Table 1)

3.2 Felt Temperature and Paralinguistic Activities with Respect to Emotions

It can be observed from Figure 4 that the people shiver and feel cold in case of fear and sadness whereas they feel warm and pleasant only in case of joy. Feeling hot and cheeks burning is maximum in case of shame and anger.

It was found that the paralinguistic symptoms (as shown in Figure 5) are the distinguishable symptoms of joy and anger in comparison with other emotions. Speech-melody change is seen maximum in case of joy and speech disturbances are seen maximum in case of anger.

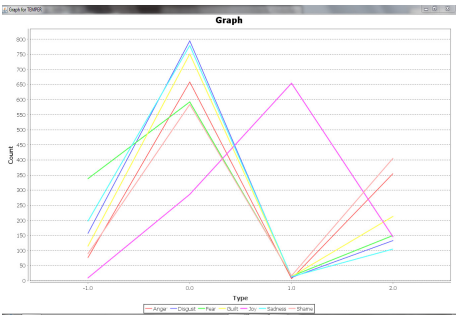


Fig. 4. Changes of Seven Emotions with respect to Different Felt Temperatures (in Table 1)

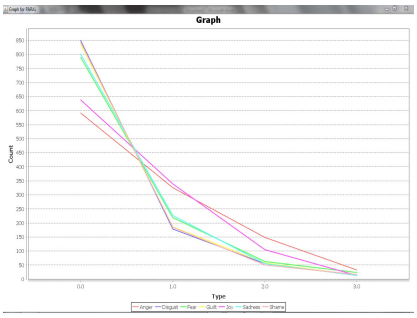


Fig. 5. Changes of Seven Emotions with respect to Different Paralinguistic Activities (in Table 1)

3.3 Movement Behavior with Respect to Gender (Male/Female) and Emotions

It has been observed that the behavior changes with respect to gender and emotions (as shown in Figure 6 and Figure 7). In case of both male and female and in case of *joy*, the *withdrawing tendency* is least and *moving towards people and things* is maximum. One interesting result for male in case of other emotions is that the *moving towards people* is more when they are *angry* whereas in female, this behavior is seen more in case of *sadness* and *fear*. But, overall, in case of both male and female, the *withdrawing nature* is maximum in case of *shame*. In case of female, the *withdrawing nature* is more in case of *disgust* than *guilt*, while it is the opposite in case of male.

Fig. 6. Changes of Seven Emotions with respect to Different Movement Behavior (in Table 1) and Gender (MALE)

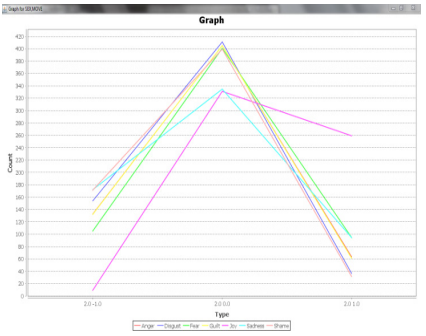


Fig. 7. Changes of Seven Emotions with respect to Different Movement Behavior (in Table 1) and Gender (FEMALE)

3.4 Verbal and Non-verbal Activities with Respect to Emotions

The symptoms of *silence* are seen maximum in case of *fear* followed by *sadness*, *shame* and *guilt*. On the other side, the *lengthy utterances* are observed maximum in

case of *joy* followed by *anger*. It is found that the people used to give *short utterances* in case of *disgust* feeling (observed in Figure 8).

In case of Nonverbal cases, *laughing and smiling* are the distinguishable symptoms for identifying *joy*. It has been observed that *crying, sobbing* and other facial expressions are changed with respect to both *joy* and *anger*. People *scream and yell* in case of *sadness*, *anger* and even in case of *joy* also (as shown in Figure 9).

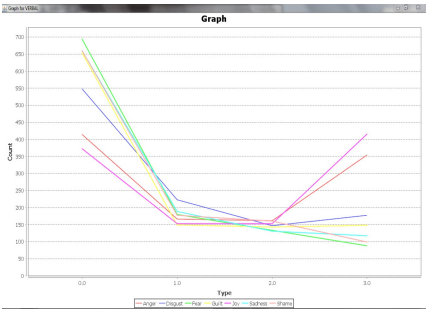


Fig. 8. Changes of Seven Emotions for Different Verbal Activities (in Table 1)

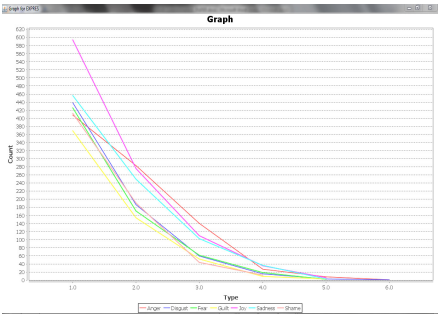


Fig. 9. Changes of Seven Emotions for to Different Nonverbal Activities (in Table 1)

3.5 Control and Self Variables with Respect to Emotions

It was observed that the people try to hide or control their feelings maximum in case of *shame* followed by *guilt* and minimum in case of *joy* followed by *anger* (as shown in Figure 10).

The positive effect on the self-esteem or self confidence of the people is very high in case of *joy*. On the other hand, the events related to *shame* and *guilt* negatively affect the self-esteem or self confidence of the people (as shown in Figure 11).

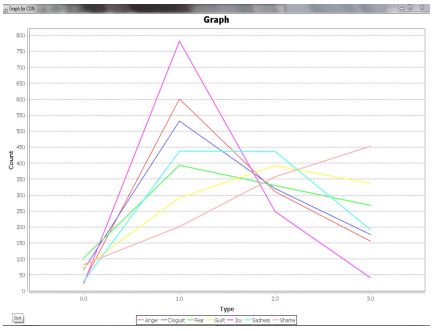


Fig. 10. Changes of Seven Emotions for Control Variables (CON) (in Table 2)

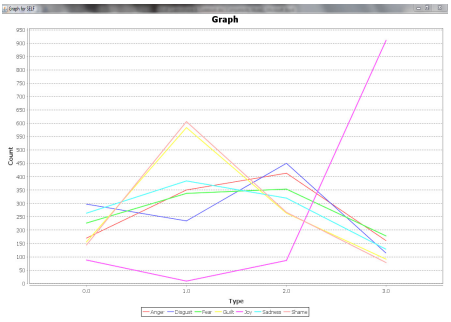


Fig. 11. Changes of Seven Emotions for Different Self Variables (SELF) (in Table 2)

3.6 Relation and Expectation Variables with Respect to Emotions

From the graph as shown in Figure 12, we can say that the nature of *relation variable* is similar to the nature of *self variable*. *Joy* also positively affects the relationships of the people involved. But, the relationships with the people involved are affected negatively in case of *anger* followed by *disgust*.

In the graph (as shown in Figure 13) of *expectation variables* and their changes with respect to emotions, it was found that the people very much expect the situation of *joy* to occur whereas they not at all expect the situation of other emotions to occur.

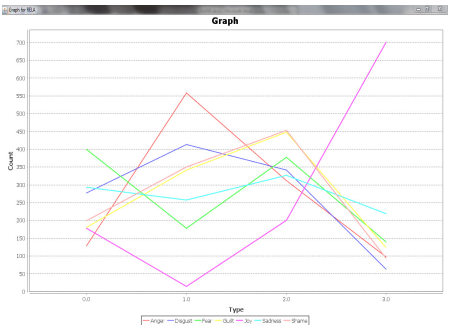


Fig. 12. Changes of Seven Emotions with respect to Different Relation Variables (RELA) (in Table 2)

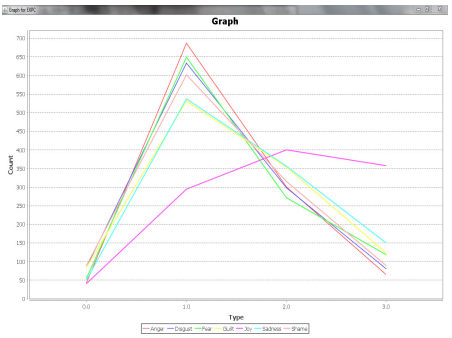


Fig. 13. Changes of Seven Emotions with respect to Different Expectation Variables (EXPC) (in Table 2)

4 Rule Based Approach to Identify Psychological Theme Words

Finally, the text clusters containing interview statements have been identified with respect to the above mentioned variables and their different combinations. But, our intention is to identify the psychological theme words present in such clusters as these theme words carry significantly useful clues to diagnose psychological symptoms in turn.

To develop a rule based baseline system to identify and classify psychological theme words, at first, we have applied a methodology for clustering the sentences (reply of the respondents of the ISEAR questionnaire) with respect to particular inputs or situations. Based on the combinations of different psychological variables, we have identified and retrieved the corresponding cluster of statements from the ISEAR database. The detailed results are described in Table 3.

After removing the stop words from the clusters, we have applied an algorithm to generate the unigram topic words based on their frequencies. *Top-10, Top-20* and *Top-30* theme words having maximum frequency are identified from the clusters with respect to that particular combination of variables (as shown in Table 4). Then, we manually collected the theme words from the clusters and compared them with the theme words generated by the system.

Table 3. Precision (Prec.), Recall (Rec.) and F-Scores (FS) and examples from Top-30 identified Theme Words for Seven Emotion Classes with respect to different combinations of Variables

Emotion Classes	Different Combinations of Variables per Class	Example Topic Words obtained from system	Based on Top-30 Topic Words		
			P	R	FS
Joy	Age in (19, 20, 21) Sex (Male) Tropho in (1, 2)	accept, passed, love, care, kiss, party, presents	.30	.64	.40
	Age in (18, 19) Sex (Female) Ergo in (2)	good, time, felt, friend, passed, happy, won, joy, glad, school	.60	.39	.33
	Age in (18, 19) Sex in (Male) Paral in (1)	accepted, received, friend, study, girlfriend, happy, snowed	.36	.39	.37
Anger	Age in (19, 20, 21) Sex in (Male) Tropho in (1, 2)	girlfriend, money, friend, angry, parents, argument, friends, school, accused	.36	.25	.29
	Age in (18, 19) Sex (Female) Ergo in (2)	money, friends, sister, angry, mother, time, friend, school, ended, promised, argument, later, authority	.40	.17	.23
	Age in (18, 19) Sex (Male) Paral in (1)	reacted, lost, friends, provoked, angry, weekend, social, time, reason, abuse, assured, carelessly, joke	.43	.72	.53
Disgust	Age in (19, 20, 21) Sex (Male) Tropho in (1, 2)	drunk, place, woman, disgusted, saw, dead, failed, friend's, cueing, fell	.40	.27	.32
	Age in (18, 19) Sex (Female) Ergo in (2)	disgusted, friends, treated, beating, girlfriend, appearance, felt, women	.36	.30	.32
	Age in (18, 19) Sex (Male) Paral in (1)	feel, money, lie, craps, bets, raped, kidnapped, disgusted	.40	.92	.55
Fear	Age in (19, 20, 21) Sex in (Male) Tropho in (1, 2)	lost, left, afraid, saw, dark, fear, night, water, road, operation, street, scared, broke	.53	.34	.41
	Age in (18, 19) Sex (Female) Ergo in (2)	scared, dark, car, afraid, father, late, night, accident, door, knocked, area, fear, man, left	.66	.39	.49
	Age in (18, 19) Sex (Female) Paral in (1)	away, fear, caught, dogs operation, angrily, reacted, afraid, waiting, barking	.30	.40	.34

Table 3. (*Continued.*)

Guilt	Age in (19, 20, 21) Sex (Male) Tropho in (1, 2)	friend, guilty, home, exam, night, father, wanted, bad, failed, caught, leave, promised	.46	.26	.33
	Age in (18, 19) Sex (Female) Ergo in (2)	guilt, friend, realized, hurt, care, lied, guilty friends, week, feeling, caught	.33	.37	.34
	Age in (18, 19) Sex (Male) Paral in (1)	belonged, caught, sneaking, forgot, missing, close, fool, refused	.33	.62	.43
Sadness	Age in (19, 20, 21) Sex (Male) Tropho in (1, 2)	passed, death, died, days, sad, relationship, girlfriend, leave, hospital, uncle, sick, friends	.50	.38	.43
	Age in (18, 19) Sex (Female) Ergo in (2)	died, felt, sad, friend, realized, suicide, passed, accident, funeral, committed, refused	.40	.40	.40
	Age in (18, 19) Sex (Male) Paral in (1)	died, relationship reason, sorrow, felt, love, exchange, deep	.26	.40	.31
Shame	Age in (19, 20, 21) Sex (Male) Tropho in (1, 2)	car, girl, caught, drunk, felt, ashamed, results, grade, thought, drink, received, started	.53	.40	.45
	Age in (18, 19) Sex (Female) Ergo in (2)	lying, asked, felt, shame, ashamed, wet, caught, party, boy, forgotten, feel, children	.36	.45	.40
	Age in (18, 19) Sex (Male) Paral in (1)	friend, class, aloud, angry, girl, forgot, hide, childish, existence, showed, reacted	.30	.42	.35

Table 4. Precision (Prec.), Recall (Rec.) and F-Scores (FS) of Seven Emotion Classes with respect to the identified Theme Words according to the rank of *Top-10*, *Top-20* and *Top-30*

Emotions	**Top-10**			**Top-20**			**Top-30**		
	Prec.	**Rec.**	**FS**	**Prec.**	**Rec.**	**FS**	**Prec.**	**Rec.**	**FS**
Joy	.56	.19	.28	.45	.30	.36	.42	.43	.42
Anger	.56	.12	.19	.41	.19	.25	.40	.27	.32
Dis.	.40	.13	.19	.33	.21	.25	.37	.36	.36
Fear	.56	.14	.22	.50	.25	.33	.50	.37	.42
Guilt	.43	.13	.19	.43	.27	.33	.37	.35	.35
Sad.	.43	.14	.21	.45	.30	.36	.40	.40	.40
Shame	.56	.20	.29	.45	.31	.36	.40	.42	.40

Different combinations of variables have been used to analyze the results of our present system. It is found that the stop words like "Didn't", "I'd", '(', ')' , -, 2, 3 were causing the errors in some of the cases. We found that the people frequently use the words like "*dark*", "*night*", "*alone*", "*left*", "*scared*", "*coming*" etc. to express the feeling of *fear* and such words alone are capable of expressing the feeling of *fear*. It is also found that these words are present as topic words with respect to various combinations of variables. But, at the same time, some other words like "*unnerved*", "*accident*" appear in case of both *sadness* and *fear* whereas the words like "*threats*", "*deserted*", "*terrible*", "*locked*", "*shadow*", "*strange*", "*assaulted*", "*pregnant*", "*horror*", "*failed*", "*broke*" etc. appear in case of *fear* only. But, all of such words are not used frequently by the people in general and for this very reason, these words are present in the clusters and do not appear in the list of topic words. As a result, in case of *fear*, the *precision* is high but the *recall* is low. As compared to other emotions, it is found that the *precision* as well as the *recall* in case of *disgust* and *guilt* are very low. The reason is that only a very few words which are capable of expressing these feelings are present in the list of topic words. The direct matches occur only for the words, "*disgust*" or "*guilt*" for identifying these classes. It is noticed that the people use the words like "*lied*", "*caught*", "*treated*", "*badly*", "*dead*" etc. to express the feeling of *disgust* as well.

5 Conclusions and Future Work

In the present work, we have developed an interface to generate visual graphs for analyzing the impacts of emotions with respect to different background, behavioral and physiological variables. We have discussed some interesting results on the visual graphs generated with respect to the variables and identified the text clusters from the interview statements by selecting individual and different combinations of such variables. The clusters are used not only for retrieving the theme words but also to classify the theme words into their respective emotion classes. In order to retrieve the psychological theme words, we have developed a rule based baseline system considering unigram based keyword spotting technique and it gives satisfactory average F-Scores for all emotion classes.

In future, we require to analyze the results by employing different machine learning techniques in the aim to achieve better results. Moreover, the textual features related to psychological variables are required to reduce the chances of errors and increase the respective F-Scores. The identification of idioms and multi-word phrases can also be carried out in our future attempts.

References

1. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of the English Language. Longman, New York (1985)
2. Grefenstette, G., Qu, Y., Shanahan, J.G., Evans, D.A.: Coupling niche browsers and affect analysis for an opinion mining application. In: Proceedings of RIAO 2004, pp. 186–194 (2004)
3. Ekman, P.: An Argument for Basic Emotions. Cognition and Emotion 6, 169–200 (1992)

4. Strapparava, C., Mihalcea, R.: Learning to Identify Emotions in Text (2008)
5. Stone, P.J.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press (1966)
6. Weintraub, W.: Verbal behavior: Adaptation and psychopathology. Springer, New York (1981)
7. Scherer, K.R.: What are emotions? And how can they be measured? Social Science Information 44(4), 693–727 (2005)
8. Das, D., Bandyopadhyay, S.: Analyzing Emotional Statements – Roles of General and Physiological Variables. In: The Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP), 5th International Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, pp. 59–67 (2011)
9. Poria, S., Gelbukh, A., Hussain, A., Das, D., Bandyopadhyay, S.: Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining. IEEE Intelligent Systems 28(2), 31–38 (2013), doi:10.1109/MIS.2013.4
10. Das, D.: Analysis and Tracking of Emotions in English and Bengali Texts: A Computational Approach. In: The Proceedings of the Ph. D. Symposium, 20th International World Wide Web Conference (WWW 2011), Hyderabad, India, pp. 343–348 (2011)

Temporal Expression Recognition in Hindi

Nitin Ramrakhiyani[1,⋆] and Prasenjit Majumder[2]

[1] Tata Research Development and Design Centre, Pune
nitin.ramrakhiyani@gmail.com
[2] DAIICT Gandhinagar
prasenjit.majumder@gmail.com

Abstract. Temporal annotation of plain text is considered as a useful component of modern information retrieval tasks. In this work, two approaches for identification and classification of temporal entities in Hindi are developed and analyzed. Firstly, a rule based approach is developed, which takes plain text as input and based on a set of hand-crafted rules, produces a tagged output with identified temporal expressions. This approach is shown to have a strict F1-measure of 0.83. In the other approach, a CRF based classifier is trained with human tagged data and is then tested on a test dataset. The trained classifier identifies the temporal expressions from plain text and further classifies them to various classes. This approach is shown to have a strict F1-measure of 0.78. In this process a reusable gold standard dataset for temporal tagging in Hindi was developed. Named the ILTIMEX2012 corpus, it consists of 300 manually tagged Hindi news documents.

1 Introduction

Presence of temporal expressions is frequent in most pieces of text. Identification and processing of this temporal information in text has been a challenging problem for the Information Retrieval and NLP communities. This identified temporal information has several beneficial applications like question answering (answering the "when" kind of questions)[1]; [2], event identification and ordering[3], and development of temporal summaries[4]; [5]. Various approaches have been developed to achieve temporal annotation of text in several European Languages(see Sect. 2). But there is a dearth of such approaches and systems for Indian Languages. In this work, analysis of two approaches for identification and classification of temporal expressions in Hindi is carried out.

The first approach utilizes hand crafted rules for identification of temporal expressions in input text. The rules are modeled as regular expressions and the input text is subjected to regular expression matching. Once a regular expression match occurs over a word or a set of words, they are tagged as a temporal expression. Further the class associated with the matched rule is assigned to the identified expression. In this work three temporal expression classes are considered namely DATE-TIME(D), PERIOD(P) and FREQUENCY(F). The rule based approach performs with a strict F1-measure of 0.83 and is presented with thorough analysis of its performance.

The other approach discussed in this work uses human tagged data to learn a classifier and then recognize temporal expressions in test text. The training text tokens are

⋆ This work was carried out when the author was pursuing Masters at DAIICT Gandhinagar.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 740–750, 2013.
© Springer International Publishing Switzerland 2013

represented as feature vectors and a CRF based classifier is trained using these training feature vectors. Next a classification is sought for vectors of test text tokens. This approach performs with a strict F1-measure of 0.78 and detailed analysis of its performance is highlighted.

Another major contribution of this work is the ILTIMEX2012 corpus that has been developed for evaluation of the discussed approaches. It contains documents with manually tagged temporal expressions and hence serves as a gold standard dataset appropriate for this task. The corpus has been developed on 300 plain documents of the FIRE 2011 Hindi Corpus[6]. This resource can be useful in a multitude of ways namely serving as a test bed for other similar systems, helping the extraction of important features of time tagged text, training a classifier to achieve the desired annotation and much more.

The rest of the paper is organized as follows. Research accomplished in the area of temporal annotation in English and other languages is reported in Section 2. A description of temporal expressions in Hindi is presented in Section 3. A formal task definition, details of the tagging and a set of illustrative examples are described in Section 4. Details about the ILTIMEX2012 corpus are presented in Section 5. Description of the rule based and CRF based approach is presented in Section 6 and 7 respectively. Evaluation results and analysis of the approaches is reported in Section 8. Finally, Section 9 presents the conclusion.

2 Motivation and Related Work

The basic motivation of this research is derived by two important facts - its usefulness in various information extraction tasks and to the best of our knowledge this being the first analysis of temporal expression identification systems in Hindi. Further the contributions of this work (systems and corpus) aim to aid contemporary research being carried out in areas of question answering, event relation and chronology development in Indian Languages. The problem also becomes challenging as there are few efficient text pre-processing systems like POS taggers and parsers available for Hindi, making it a poor-resourced language.

A good amount of successful research has been accomplished in temporal expression annotation in languages namely English[7], Italian[8], Spanish[9], German [10] and in Chinese[11]. Starting from the work on the TIMEX at the Message Understanding Conference[12] till recently on approaches like Time Aware Information Access(TAIA) [13], there have been developed and deployed, several temporal annotation systems. A pool of such resources is available at TIMEX Portal[14]. Two major contributions in this area are the temporal annotation guidelines for English, developed and published by MITRE Corporation, known as the TIMEX2 standards - year 2001[15] and year 2005[16]. These guidelines present a lucid framework of differentiating temporal expressions from other words in text and normalizing the temporal expressions. At the TERQAS 2002[1] workshop, a new markup language, TimeML for temporal annotation was contributed by [2]. TimeML aimed at identifying events, temporal expressions and temporal relations between these events. It gave definitions of several tags namely TIMEX3, EVENT, SIGNAL, TIMEML to address the complex event and relation identification.

Mani et. all. in [7] focuses specifically on a multilingual approach to temporal annotation and explains a simple technique to parallel the English temporal tagging program for other languages. In their work, the developed time tagger system is shown to have an F-measure of 0.962 with identification and 0.832 when considering values of identified expressions.

Another system, Chronos[8] was developed at the ITC-irst for the ACE TERN 2004 evaluation. This system applies the TIMEX2 tagging to Italian and English texts and was shown to have an F-measure of 0.926 with identification and 0.872 when considering values of identified expressions. Another system which generates TIMEX2 annotations in English text is known as the DANTE[17]. It was developed at Maquire University and highlights an intermediate semantic representation of temporal expressions which is in line with the TIMEX2 standards.

HeidelTime[10], developed at University of Heidelberg works using a rule base and employs regular expression matching for extraction and normalization of temporal expressions. HeidelTime is shown to have an F1-measure of 0.86 in temporal expression identification.

The above mentioned rule based systems use a large set of hand crafted rules. To minimize this dependence on a rule set, a machine learning approach was proposed by Ahn et. all.[18]. This new architecture replaced the rule base by a set of machine learned classifiers to achieve the desired temporal expression annotation in English. Their system, TimexTag is shown to have an F-measure of greater than 0.8. Another system developed at CSLR, University of Colorado[11] uses a feature based method for temporal expression extraction in English and Chinese text. The system performs classification in a token-by-token manner and their feature set is supported by cues from a rule based and a statistical tagger for the extraction task.

3 Temporal Expressions in Hindi

A temporal expression in Hindi is formed of one or more words which collectively represent a point or a duration or frequency in time. Known and widely used Hindi time words like date and time formats, names of days, name of months, name of seasons, etc form what are known as *pivot words*. A Hindi temporal expression is assumed to contain one or more of these pivot words. Words which quantify or modify the pivot word are also considered a part of a temporal expression. They are known as quantifiers. Direction words give a time direction to the pivot word and are also considered as a part of a temporal expression. Table 1 presents a representative set of pivot words, quantifiers, direction words and other temporal expression constituent words.

4 Task Definition

The aim of the task is to achieve the two following goals.

Identification of temporal expression in given input text: To fix a boundary around one or more words which denote(s) a proper temporal expression.

Table 1. Markable Temporal Expressions

Type	Words
Always Marked	
Pivot Words - Nouns	युग(yug,“epoch”), सदी(sadi,“century”), वर्ष(varsh,“year”), दिन(din, “day”)
Pivot Words - Day and Month Names	सोमवार(somvar,“Monday”), मंगलवार(mangalvar, “Tuesday”), जनवरी (janvari,“January”), फरवरी(farvari,“February”)
Pivot Words - Known Date and Time Formats	०१-जून-२०१०(01-Jun-2010), ०१/जून/२०१०(01-Jun-2010), १०:३०(10:30), १०.३०(10.30)
Present, Past and Future	भूतकाल(bhootkal,“past”), भविष्यकाल(bhavishyakal, “future”), वर्तमानकाल (vartamankal, “present”)
Festivals and Seasons	दिवाली(diwali, “Diwali”), होली(holi, “Holi”), गरमीयां(garmiyan, “Summers”), सरदियां(sardiyan, “Winters”)
Marked as part of temporal expression if they quantify a pivot word	
Quantifiers	चंद(chand,“few”), कुछ(kuch,“some”), थोड़े(thode,“little”), बहुत(bahut,“more”)
Hindi Numerals and Number Words	एक(ek,“one”), दो(do,“two”), तीन(teen,“three”), १(1),२(2),३(3)
Marked as part of temporal expression if they direct a pivot word	
Direction Words	अगले(agle,“next”), पिछले(pichle,“last”), पहले(pehle,“before”), बाद(baad,“after”)
Marked as part of pre or post modifiers	
Other Words	को(ko,“at”/“to”), के(ke,“on”), का(ka,“of”), की(kii,“of”), में(mein,“in”), से(se,“from”), करीब(kareeb,“about”)

Classifying the identified temporal expression: To classify the identified temporal expression as one of the three temporal expression classes - PERIOD(P) denoting a time period or duration, DATE-TIME(D) denoting a date or time point and FREQUENCY(F) denoting a time frequency.

To carry out the delimitation of a temporal expression and for assigning it an attribute (target classification), tagging the required set of words within an XML tag is an appropriate method. In the current experiments a single XML tag: <TIMEX> is used to delimit the set of words forming a temporal expression. This tag is devised to have a single attribute namely TYPE. This attribute holds a value denoting the class of the enclosed temporal expression namely P, D or F for PERIOD, DATE-TIME OR FREQUENCY respectively. To highlight the desired tagging, following are some sample tagged sentences.

Plain Input: मैं हफ्ते भर का सामान वहीं से लेता हुँ।(main hafte bhar ka saman vahin se leta hu)
TER Output: मैं <TIMEX TYPE="P">हफ्ते भर</TIMEX> का सामान वहीं से लेता हुँ।(Time Entity: हफ्ते भर (hafte bhar, "week long"))

Plain Input: कॉलेज में १४ जून २०१० को परिणाम आयेगा।(college mein 14 Jun 2010 ko parinam aayega)
TER Output: कॉलेज में <TIMEX TYPE="D">१४ जून २०१०</TIMEX> को परिणाम आयेगा।(Time Entity: १४ जून २०१० (14 Jun 2010))

Plain Input: हमारा मुनाफा सालाना छह से सात प्रतिशत बढ़ता है।(hamara munafa saalana chah se saat pratishat badta hai)
TER Output: हमारा मुनाफा <TIMEX TYPE="F">सालाना</TIMEX> छह से सात प्रतिशत बढ़ता है।(Time Entity: सालाना (saalana, "yearly"))

5 The Document Collection

The ACE TERN 2004 Evaluation Corpus[19] was released under the Automatic Content Extraction(ACE) Experiment[20] of 2004. It was aimed at evaluation of Entity Detection-Recognition, TIMEX2 Detection-Recognition and other tasks. It consisted of about 50K English words collected from Broadcast News programs and Newswire reports. To evaluate the performance of the developed systems similar gold standard data is necessary. However, to the best of our knowledge, there is presently no temporal expression tagged corpus in Indian Languages. So an important task was to prepare such a corpus for evaluation.

The FIRE 2011 Hindi corpus[6], which has a rich collection of news articles from the Hindi newspaper Navbharat Times, was considered suitable. A subset of documents, each having more than 500 words, was collected from the FIRE 2011 Hindi Corpus. Manual temporal tagging of the chosen documents would only provide the necessary quality for consideration as gold standard data. Hence, for the development of the test bed, manual tagging of the chosen documents was carried out. This manual tagging was carried out using the General Architecture for Text Engineering(GATE) Tool[21]. The set of plain documents along with their corresponding manual tagged documents is collectively named as the ILTIMEX2012 corpus. Table 2 highlights a comparison between the ILTIMEX2012 Corpus and the ACE 2004 TERN English Evaluation corpus.

Table 2. Document Collection Statistics

Detail	**TERN 2004 Eval Corpus**	**ILTIMEX2012**
Total number of Documents	192	300
Nature of Documents	News	News
Total number of words	About 50K	About 150K
Total number of temporal expressions	1828	1840

6 The Rule Based Approach

In this approach, a set of manually developed rules is employed to identify and classify temporal expressions in plain text. The realization of the rules is done through Regular Expressions(REGEX). This eases the process of handling a large number of rules and makes the identification process effective. For example, for separate rules to identify expressions like १४-०८-२०१०(14-08-2010), १४/०८/२०१०(14/08/2010), १४.०८.२०१०(14.08.2010) a single regular expression - ([०१२३][०१२३४५६७८९])[-/.]([०१][०१२३४५६७८९])[-/.]([०१२३४५६७८९]{4}) is included. English representation of this expression is ([0123][0123456789])[-/.]([01][0123456789])[-/.]([0123456789]{4}). An initial set of rules is first compiled with pivot words, quantifiers and direction words representing valid tag extents as per the decided tagging guidelines. Also a class value(P, D or F) is stated with each rule, denoting the class of temporal expressions it aims to identify. After application of rules are to texts in training data, the feedback from the output is used to iteratively enhance and improve the rule base.

Linguistic preprocessing like POS tagging or chunking of the input plain text is not carried out. The current approach relies entirely on the REGEX based rules for identification and classification. The application of rules starts by matching each of the REGEXs in the initial rule base through the input text. If a REGEX match occurs, the matched collection of words are placed under the <TIMEX></TIMEX> tags signifying a temporal expression. The TYPE attribute of the tag is set to the class value of the matched rule.

The ILTIMEX2012 corpus was divided into 200 documents for iterative rule base development and 100 documents for testing the developed rule base. The initial rule base was compiled and the training files were processed using it. The results were compared with the corresponding manually tagged files. This comparison let to identification of partially correct, missing and false positive entries. An analysis of these entries helped in modifying the rule base by addition of new rules and modification of old rules. Figure 1 depicts a graph of addition of new rules against number of training documents analyzed. The graphs is observed to flatten at the end indicating a convergence for the set of documents considered.

The number of rules in the current system stands at 69, which is small with respect to the large number of different temporal expressions that get identified using the rule base. The rule base continues to be enhanced by addition of new rules. As an example of the rules, following is presented, the rule that is used the most number of times while performing evaluation. The type names given in Table 1 are used to explain the rule.

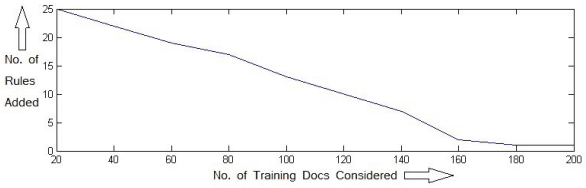


Fig. 1. Rule Base Enhancement

Rule: (Pivot Words - Nouns)(\s(Other Words के(ke) and से(se))\s(Direction Words))?
Examples: वर्ष(varsh, "year"), दिन(din, "day"), पल(pal, "second"), साल से पहले (saal se pehle, "year before")
Usage Frequency: 112

7 The CRF Based Approach

A rule based approach always requires a set of hand crafted rules which can be a costly resource if number of iterations for manual addition of rules increase. Porting such a rule base to another language can further prove to be tedious. Prevalent practice and literature suggests another well known approach that involves learning a classifier to perform an entity recognition task.

This approach involves representation of human annotated data in form of feature vectors and training a classification system using these feature vectors. Once the classifier is sufficiently trained, it can be tested on plain text to perform the desired identification and classification of temporal expressions. Two important requirements for this approach to succeed are identification of proper features and secondly, proper representation of training data words in terms of identified features.

In the current experiment we use an implementation of a Conditional Random Field (CRF) based classifier. Standard CRF implementations work with two kind of features namely token level and context level. A set of 4 token level and 20 context level features was identified for this task. The token level features are briefly described as follows.

- **EngRep:** Representation of Hindi word in English. The transliteration oriented, WX format representation of Hindi text in English was used. The WX representation is a ASCII based representation explained in [22].
- **POSInfo** - POS tag information of the token. The POS tagger provided by TDIL, India and developed at IIIT, Hyderabad was employed in the experiment.
- **isPivot** - Is 1 if the token is a pivot word, otherwise 0. This is a list based feature which checks if the token is a pivot word.
- **containsDigitsAndPunc** - Is 1 if the token contains digits and punctuation marks, otherwise 0. This feature checks for tokens with numbers, punctuations and with both numbers and punctuations. Through this feature numeric tokens quantifying pivot words can be captured.

A representative set of context level features are described as follows. The context window is considered as ± 2 tokens from the token under consideration.

- **POSInfo**($i \pm 2$)**, POSInfo**($i \pm 1$)**, POSInfo**($i \pm 0$) - POS feature information for tokens in context window around token i.
- **EngRep**($i \pm 1$)**, EngRep**($i \pm 0$) - English Representation for tokens in context window around token i.
- **containsDigitsAndPunc**($i \pm 1$)**, isPivot**($i \pm 0$) - isPivot of token i and containsDigitsAndPunc of token ($i \pm 1$).

The training data also provides information about the target class which is the basic learning point for the classifier. The representation of this class information is carried

out using the BIO model where the starting token of the tag is signaled as B, intermediate tokens of the tag as I and all tokens outside the tags as O. Further the single character TYPE information i.e. D, E and F is appended to the B and I symbols using a hyphen. This provides a complete representation of the tag boundary and TYPE information.

As in the rule based approach, a division of 200 files for training and 100 for testing is used. The training data is converted to a token-features format containing a single token and the associated token level features on a single line. The context level features are described in a template file provided to the CRF classifier while training. The token-features format training data and the context feature template file is supplied to the CRF classifier to learn a model. The set of test documents are then tagged using the learned model.

8 Evaluation

The systems developed based on the discussed approaches, were made to tag the plain testing documents and the output files were compared with the corresponding manually annotated files. For the performance evaluation, the Annotation Difference Tool of the GATE tools[21] was employed. Benefits of the GATE annotation difference tool are its visual comparison interface, file editing before comparison and automatic metric value computation. Tables 3 highlights the obtained experimental results from the discussed approaches.

Table 3. Evaluation - Rule Based and CRF Approaches

	Rule Based		CRF	
	Id	Id + Cl	Id	Id + Cl
Manual Tagged	650	650	650	650
Total Tagged	668	668	639	639
Correct	593	544	543	502
Partially Correct	38	23	67	44
Missing	19	83	40	104
False Positive	37	101	29	93
Strict Evaluation				
Recall	0.91	0.84	0.84	0.77
Precision	0.89	0.81	0.85	0.79
F1-measure	0.9	**0.83**	0.84	**0.78**
Lenient Evaluation				
Recall	0.97	0.87	0.94	0.84
Precision	0.95	0.85	0.96	0.85
F1-measure	0.96	**0.86**	0.95	**0.85**

Id: Identification; Id + Cl: Identification and Classification

Lenient Evaluation considers Partially Correct as Correct

8.1 Analysis

Analysis of evaluation of the developed systems is critical and is presented as follows. The focus of the analysis is on the missing entries observed during the evaluation. Missing entries get generated when a temporal expression is completely missed by the system or when the system identifies the expression with right boundaries but fails to assign it the right TYPE. Temporal expressions missed individually by the systems and by both together are highlighted below.

Expressions missed by the Rule Based system

- Expressions denoting a year value like 2012, 1998, etc are missed at times. The rule base consists of a rule to tag 4-digit words as temporal expressions when they are followed by prepositional post modifiers like में(mein,"in"), से(se,"from"), etc. When these 4-digit year expressions occur without such post modifiers the expression is missed.
- Expressions not recognized by any rule. These expressions are not included in the rule base due to absence of such expressions in training data and hence during iterative rule base development.

Expressions missed by the CRF Based system

- Expressions of the form दिनभर(dinbhar,"day long") are missed. Expressions of this form are frequently expressed with a space between the two words "दिन"(din,"day") and "भर"(bhar,"long") and less frequently without the space. The classifier gets trained accordingly and hence fails when the separating space is not present.
- Expressions that occurred less frequently in the training data. This is a very intuitive reason for certain expressions being missed by CRF while testing. Example expressions are समय-समय(samay-samay,"time-to-time"), सौ डेढ़-सौ साल(sau dedh-sau saal, "100 to 150 years")

Expressions missed by both the systems

- Conjuncted parts of a single temporal expression are tagged separately as per the guidelines. Both the approaches fail to identify the pivot less parts of the expression. For example, if the desired expression tagging is <TIMEX TYPE="D">1</TIMEX> और <TIMEX TYPE="D">2 जनवरी</TIMEX> the pivot less portion i.e. <TIMEX TYPE="D">1</TIMEX> is always missed and remains untagged.
- Temporal pivot words in English which are transliterated and used in Hindi are not identified by both the approaches. For example सन्डे("Sunday") and सेकंड("second").
- Specially expressions like "9/11" are missed by both the systems. The rule base has no rule to identify this kind of a numeral-punctuation combination as a temporal expression. An inclusion of such a rule would lead to tagging of similar non-temporal tokens which are much frequent than these special expressions. The CRF approach misses it due to absence of such expressions in training data.

The GATE Tool evaluation also considers expressions which have been identified correctly with the right tag boundaries but incorrect attribute value as missing entries. As a resut, the number of missing entries jump from a low 19 to a high 83 when the TYPE attribute gets considered for evaluation. Similarly for the CRF based approach the missing entries jump from 40 to 104. It is also observed that in more than 98% cases of misclassification, the types involved are P(PERIOD) and D(DATE-TIME), and is a major challenge to the described approaches.

9 Conclusion

The paper describes two approaches to temporal expression identification and classification in Hindi - a rule base approach (F1-measure: 0.83) and a CRF based approach (F1-measure: 0.78). Evaluation analysis of both approaches is elaborated with examples. As future work, the approaches will be fused to combine merits of both.

Acknowledgments. This research is supported partly by the *Cross Lingual Information Access* project funded by *D.I.T., Government of India*. The authors would like to thank IRLAB DAIICT member Harsha Kokel for her support in manually tagging the Hindi Documents.

References

1. Pustejovsky, J.: TERQAS: Time and Event recognition for question answering systems. In: ARDA Workshop (2002)
2. Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: TimeML: Robust specification of event and temporal expressions in text. In: New Directions in Question Answering 2003, pp. 28–34 (2003)
3. Mani, I., Schiffman, B.: Temporally anchoring and ordering events in news. In: Time and Event Recognition in Natural Language. John Benjamins (2005)
4. Swan, R., Allan, J.: Automatic generation of overview timelines. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 49–56. ACM (2000)
5. Allan, J., Gupta, R., Khandelwal, V.: Temporal summaries of new topics. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 10–18. ACM (2001)
6. Majumder, P.: Forum for Information Retrieval Evaluation 2011 (2011), `http://www.isical.ac.in/ fire/2011/slides/ fire.2011.majumder.prasenjit.pdf`
7. Mani, I., Wilson, G., Sundheim, B., Ferro, L.: A multilingual approach to annotating and extracting temporal information. In: Proceedings of the Workshop on Temporal and Spatial Information Processing, vol. 1, p. 12. Association for Computational Linguistics (2001)
8. Negri, M., Marseglia, L.: Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Rapport Interne, ITC-irst, Trento (2004)
9. Saquete, E., Muñoz, R., Martínez-Barco, P.: Event ordering using TERSEO system. Data & Knowledge Engineering 58(1), 70–89 (2006)

10. Strötgen, J., Gertz, M.: HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324 (2010)
11. Hacioglu, K., Chen, Y., Douglas, B.: Automatic time expression labeling for english and chinese text. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 548–559. Springer, Heidelberg (2005)
12. MUC-7: Message Understanding Conference 1998. In: Proceedings of the Seventh Message Understanding Conference, DARPA (1998)
13. Shokouhi, M.: SIGIR 2012 Workshop on Time Aware Information Access (2012), http://research.microsoft.com/en-us/people/milads/taia2012.aspx
14. Mazur, P.: TIMEX Portal (2008), http://www.timexportal.info/
15. MITRE-Corporation: TIDES Temporal Annotation Guide. The MITRE Corporation (June 2001)
16. MITRE-Corporation: 2005 Standard for the Annotation of Temporal Expressions. The MITRE Corporation (April 2005)
17. Mazur, P., Dale, R.: An intermediate representation for the interpretation of temporal expressions. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions, pp. 33–36. Association for Computational Linguistics (2006)
18. Ahn, D., Rantwijk, J., Rijke, M.: A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In: HLT-NAACL, pp. 420–427 (2007)
19. NIST: The ACE 2004 Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/ace/2004/doc/ace04-evalplan-v7.pdf (2004)
20. NIST: Automatic Content Extraction (2004), http://www.itl.nist.gov/iad/mig/tests/ace/2004/index.html
21. Cunningham, H.: GATE, a general architecture for text engineering. Computers and the Humanities 36(2), 223–254 (2002)
22. Bharati, A., Chaitanya, V., Sangal, R., Ramakrishnamacharyulu, K.: Natural language processing: a Paninian perspective. Prentice-Hall of India, New Delhi (1995)
23. Jha, G.N.: The TDIL Program and the Indian Langauge Corpora Intitiative(ILCI). In: LREC (2010)
24. Kodu, T.: CRF++: Yet another CRF toolkit (2005), http://crfpp.googlecode.com/svn/trunk/doc/index.html

A Joint Source Channel Model for the English to Bengali Back Transliteration

Tirthankar Dasgupta, Manjira Sinha, and Anupam Basu

Indian Institute of Technology Kharagpur
tirtha@cse.iitkgp.ernet.in

Abstract. In this paper we present an English-to-Bengali back transliteration system that can be used to transliterate Bengali texts written in Romanized English, back to its original script. Our proposed system uses a bilingual parallel corpus of English-Bengali transliterated word pairs and applies both the orthographic as well as phonetic information to two different computational models namely, the joint source channel model and the trigram model, to automatically identify, extract and learning of transliteration unit (TU) pairs from both the source and target language words. Finally, the system predicts the top 10 best possible outcome of the given input text. We further extend our work to make the target word prediction module more robust. This is done by the phonological analysis of the generated target sentence. Both the models have been evaluated with a set of 2000 Romanized Bengali test words. Our initial evaluation results clearly shows that the joint source channel model performs much better than the trigram model.

Keywords: Bengali transliteration, joint source channel model, evaluation.

1 Introduction

Transliteration is a process of phonetically translating words of one language (L_1) into the same word in another language (L_2) where, alphabets ($\sum_1$*) of L_1 are expressed using alphabets ($\sum_2$*) of L_2 maintaining the same pronunciation of L_1 (Knight and Graehl, 1998). This approach is commonly known as forward transliteration. Similarly, back transliteration is the exact opposite of the forward transliteration process where words of L_2 that are expressed with $\sum_1$*of L_1 are transliterated back into the original language L_2 with alphabets $\sum_2$*. As an example, for English (L_1)-Bengali (L_2) language pair, transliteration of the word "কম্প্যুটা (kamapYuTAra)" to "computer" is considered to be a forward transliteration whereas, "amra" to "আমরা (AmarA)" is the back transliteration process.

The process of transliteration has got numerous in the field of natural language processing like, machine translation, cross language information retrieval and information extraction. Further, there has been a growing interest of using back-transliteration techniques to transliterate target language texts written in Romanized scripts to the actual target language scripts such as English to Bengali, Hindi or any other Indian language texts. This is particularly useful for countries like India where most of the existing keyboards support only English language characters. As a result it becomes difficult for a person to type Bengali texts using the English keyboard

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 751–760, 2013.
© Springer International Publishing Switzerland 2013

layout. Thus, building a back transliteration system that can automatically transliterate any Indian language texts written in roman English to their original Indian language script will be utterly useful.

Plethora of works have been done to develop computational models for back transliteration, however, very few works have been done in the English-Bengali back transliteration. Therefore, in this paper we present an English-to-Bengali back transliteration system that can be used to transliterate Bengali texts written in Romanized English, back to its original script. The primary reason behind choosing Bengali as our target language is the fact that Bengali is considered to be the second largest language spoken in India and it is the seventh largest language spoken all over the world.

Our proposed system uses a bilingual parallel corpus of English-Bengali transliterated word pairs and applies the joint source channel model to automatically identify, extract and learning of transliteration unit (TU) pairs from both the source and target language words. Although joint source channel model and its different variants have already been applied to English-Bengali back transliteration, however, the primary contribution of this paper is to propose a novel way of creating a bilingual parallel corpus that can be used to train the joint source channel model and finally evaluating the models with the matrices that have not been used previously to evaluate English-Bengali transliteration models. The proposed system predicts the top 10 best possible outcome of the given input text. We further extend our work to make the target word prediction module more robust. This is done by the phonological analysis of the generated target sentence.

The rest of the paper is organized as follows: Section 2 discuss about the various issues related to transliteration, section 3 discuss about some of the existing models on transliteration and back transliteration. In section 4 and 5 we present joint source channel transliteration model, rules for identifying TU pairs and learning of the alignment rules. In section 6 and 7 we present our evaluation technique and results of evaluation. Finally, in section 8, we conclude the paper.

2 Issues Related to English-Bengali Transliteration

Transliteration is an inherently ambiguous process. The problem of back-transliteration is generally considered to be much more ambiguous than the forward transliteration. This can be attributed by the following reasons:

1. Depending upon the context, a single English alphabet may have multiple Bengali representations. For example, "d" may be represented in Bengali as "দ", "ধ", "ড়", "ঢ়", "ড", and "ঢ". Further, a single Bengali character may have multiple English representations. For example, the character "ড়" in বাড়ি (bADi) can be transcribed as "r" (as in "*bari*") or "d" (as in "*badi*"). This phenomenon is illustrated in Fig. 1 below.

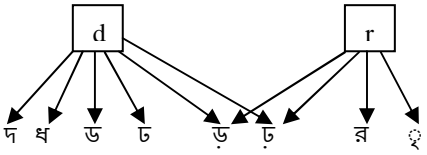


Fig. 1. Ambiguities in Bengali Transliterations

2. Multiple source language representation may map to a single target language word. For example, "*bhalo*" and "*valo"* maps to the same word "ভাল (bhAla)".
3. Two or more source language words can map to a single target language word thus, creating ambiguities among the target language words. For example, consider the source and target languages to be English and Bengali. The English input "*amar*" may map to the target language "আমার (AmAra)" or "অমর (amara)".
4. Due to the partial phonetic nature of Indian language scripts, it is often observed that the pronunciation of a word does not match the way it is written. For example, the word "নক্ষত্র (nakShatra)" is pronounced as "*nOk*। *khot*। *ro*" where, the conjugate ক্ষ (*kSha)* is pronounced as ক্খ (*kkha)*. Similarly, ক্ষ্ম (kshma) in "সূক্ষ্ম (sukShma)" is pronounced as "*kkha*". However, the word "ক্ষত্রিয় (kShatriYa)" is pronounced as "*khOt*। *ri* ।*Yo*" here, the conjugate ক্ষ (*kSha)* is replaced by খ (*kha)*.
5. Another interesting phenomena is commonly observed in Bengali where pronunciation of a character sequence is just the opposite of the way it is written as in the case of " চিহ্ন(chihna)" which is pronounced as "*chin*। *ho*". Here, হ্ন (hna) is pronounced in as ন্হ (nha).

3 State of the Art

Lee et al (1998) has proposed an English-to-Korean transliteration system. The model is a grapheme based model. Every input source English word is segmented into different pronunciation unit; a pronunciation unit is defined as a grapheme that corresponds to the source phoneme. In the next step the model assigns the most appropriate Korean grapheme corresponding to each pronunciation unit in the English word.

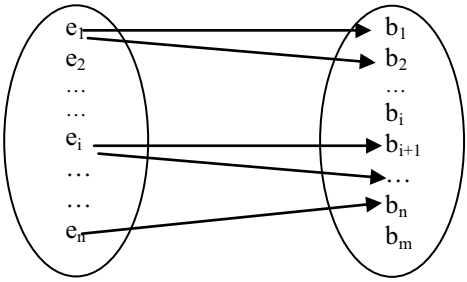
Goto et al (2003) has put forward a syllable based English to Japanese transliteration scheme. A conversation unit is a partial character strings in an English word. Each English word is divided into different conversation units and each conversation unit is aligned with a partial Japanese Katanaka character string. The likelihood of a particular chunking set for an English word is done by a syllable based correspondence between English and Katanaka characters. The model employs a maximum entropy model to handle the cases where an English conversion unit can be mapped to more than one Japanese conversion units.

A framework for direct orthographical mapping (DOM) between a language pair has been proposed by Haizhoou et al. (2004). The model uses a joint source channel model that is termed as n-gramtransliteration model. The orthographic alignment process is automated and aligned transliteration units are derived from a bilingual dictionary. The system has been tested for English-Chinese pair and found to perform better than the state-of-the-art.

4 The Joint Source Channel Model

The joint source-channel transliteration model (Haizhou et al., 2004) (n-gram TM) tries to capture the source-target word orthographical mapping relation and the contextual information. Unlike applying the approach to map the source word with the

target word, the joint source channel model tries to generate the target word from the given source word orthographical representation. Let us consider the source English language word be E =e_1 e_2 e_3 ….e_n and the target Bengali word B = {b_1 b_2 b_3…b_m} where m≠n. Here, ei and bi (1≤ i ≤ n) are English and Bengali alphabets. Both the English as well as the Bengali words can be divided into multiple substrings which are defined as transliteration unit (TU) (Ekbal et al. 2006). Each English TU ei can be aligned to that of the corresponding Bengali TU bi to form a transliteration unit pair denoted as <ei,bi>. Thus the alignment Y between E and B is denoted as <e, b>$_1$ = <e1, b1>, <e, b>$_2$ = <e2 , b2>, <e, b>$_k$ = <e_k , b_k>. However, it is to be noted that depending upon the TU definition the TU alignment between E and B is not always one to one i.e it is possible to have a single English TU (ETU) to have more than one Bengali TU (BTU).



From the above definition, the English to Bengali transliteration is defined as

$$E = \arg\max_{E,Y} P(E,B,Y) \tag{1}$$

Based on equation (1) the n-gram transliteration model for English to Bengali language pair is defined as the conditional probability of the alignment Y for the transliteration pair <e,b>$_j$ depending upon its previous n predecessor pairs. This is expressed as:

$$\begin{aligned} P(E,B,Y) &= P(E,B|Y) \times P(Y) \\ &= \prod_{k=1}^{k} P\left(< e,b >_k \,\middle|\, < e_{i-1} >_1^{k-1}\right) \end{aligned} \tag{2}$$

5 Back Transliteration of English to Bengali

To overcome the limitations of phoneme-based approach our proposed English to Bengali machine transliteration system applies the joint source channel model for the direct orthographical mapping of source and target language words, as discussed in (Goto et al., 2003; Haizhou et al., 2004). However, unlike using probabilistic model for identification of TUs, we prefer to follow the techniques proposed by (Ekbal et al. 2006) where, TU identification is done using some predefined language specific linguistic rules. The rules identify both phonetic as well as grapheme information from both the source as well as target language word. The phonetic information is identified using the grapheme to phoneme conversion tool. Further, the TU identification rules use contextual information in like, diphthongs, semivowels and conjuncts. In the following sub sections we will briefly discuss about building the bilingual corpus to

perform source-target TU alignments, TU identification and alignment rules, training and the testing phase.

5.1 The Bilingual Training Corpus

As discussed above grapheme based transliteration approach requires a large bilingual parallel corpus. The corpus is used to automatically extract the source and target language TU pairs using some predefined rules. For resource poor languages like Bengali, building such a large corpus is difficult and thus it is not available so far. Hence under the current scenario we propose a novel way of automatically generation a bilingual parallel corpus for the English-Bengali machine transliteration.

Bengali being a part of Indo-Aryan script is considered to be partially phonemic in nature. Partially phonetic languages are those that use writing systems which are in between strictly phonemic and non-phonemic. Most of the Modern Indo-Aryan languages (MIA), being derived from Sanskrit, are partially phonemic. They still use Sanskrit orthography although the sounds and pronunciation rules have changed to varying degrees.

The training corpus contains around 83000 English Bengali bilingual data. This data is created from three different sources. First, we have collected the most frequent 5000 words from the Anandabazar corpus[1]. These 5000 words are manually transcribed by four different volunteers. The reason behind using multiple volunteers is to capture the diversity in representing a single word by different users as discussed in section 2. Further, we have chosen the 5000 Bengali words such that the wordlist contains all possible Bengali consonant clusters, vowel-vowel combinations (i.e. the diphthongs) and the consonant-vowel combinations. The next part of the data is a collection of manually transcribed 8000 proper nouns that consist of names of person, and places. The third part of the corpus is a collection of 70000 common Bengali content words. These words are phonetically analyzed and their pronunciations are used for the training purpose. The phonetic analysis of the common Bengali words is performed using the grapheme to phoneme converter (G2P). The G2P engine takes a Bengali word in iTrans notation[2], and returns the proper pronunciation of the input word identifying the syllables, and removing the schwas occurring in a word.

5.2 Identification of TU Alignment Rules

Identification of transliteration units is not a trivial problem. It essentially involves two important steps. First, identification of both source and target language TU and second, alignments of the source and target TUs. As discussed in (Ekbal et al., 2006) we define the English transliteration units as the character sequence C*V* where C* is the sequence of consonants and V* is the sequence of vowels. Apart from the standard set of English vowels a, e, i, o, u we have also considered the two semi vowels w and y as vowels. Unlike the English TU, the Bengali TUs are defined in terms of Akshars. For Indian language scripts, an Akshar is defined as a sequence of

[1] Downloaded from
www.cel.itkgp.ernet.in/~resources/anandabazar_corpus/

[2] http://www.cs.cmu.edu/~madhavi/Om/Mapping.html

character strings that consist of either a single consonant(C), a consonant cluster (CC or CCC), consonant followed by vowel (CV) or a consonant cluster followed by vowel (CCV or CCCV). We redefine the Bengali vowel set of 11 vowels by considering the consonants ৎ (.ta), ্য)jPhala) and য় (Y) as vowels. Each of the source language TU is mapped to the corresponding target language TU thus forming a source-target TU pair. We illustrate this with an example, suppose the English word E is "sibansu" and the corresponding Bengali representation B is "সিবাংশু) sibA~Nsu)". We can segment "sibansu" as "si | ba| nsu" and the Bengali representation as "সি)si) |বা (bA) |ংসু (~Nsu)". The alignment Y is defined as <si,সি>, <ba, বা>, <~Nsu, ংসু>. However, as discussed in section 4 it is not necessary that both the source and target word will have the same number of TUs. It is possible to have a single English TU (ETU) to have more than one Bengali TU (BTU). For example, "Australia" in English, has only three TUs "Au |stra |lia" whereas in Bengali অস্ট্রেলিয়া has four TUs অ |স্ট্রে |লি |য়া. Under such circumstances we apply some rules to identify the correct TU alignments. For example, to handle the above case we construct a list of diphthongs to identify the vowel-vowel clusters. Using this list the system can easily identify িয়া as a diphthong and thus it will consider লিয়া as a single TU rather than two different units. Again, for words like "খাই (khAI)" ,"লাউ(lAU) ,"and" খায়া(khAYa)", we have prepared a separate list of semi vowels that are used to identify and align the English and Bengali TUs. Apart from the occurrence of semi-vowels and diphthongs there are certain words for which the ETU-BTU alignments are difficult. For example, the Bengali word "আচমকা) AchamakA)" is written in English as "achomka". Based on our definition of the ETU and BTU, "achomka" is segmented into three TUs "a| cho| mka" on the other hand "আচমকা)AchamakA)" is segmented into four TUs "আ(A)| চ(cho) | ম(ma) | কা(kA)" here we can observe that the ETU "mka" is incorrectly aligned to "ম(ma)". We handle such situation by preparing an exhaustive list of consonant-consonant clusters. From this list we can identify that "ম (ma)"and "ক (ka)" does not form any Bengali conjugate and so the ETU "mka" will be further segmented as "m| ka".

5.3 The Training Phase

The English-Bengali back transliteration system is essentially composed of three major steps.

- Identification of the TUs from both the source as well as target language words.
- Using a bilingual corpus to automatically learn the mapping of the source language TU with the corresponding target language TU.
- Identify the target language TUs which returns in the highest probabilistic result.

The machine transliterated Bengali word is obtained using direct orthographic mapping by identifying the equivalent Bengali TU for each input English TU and finally placing the English TUs in order. For this, a decoder has been built that takes the source TUs and the generated target language TUs, and searches for the best probabilistic path of the given transliteration pairs by resolving different combinations of TU alignments. Similar to the approach taken by (Haizhou et al., 2004) we used the stack decoder (Schwartz et al., 1990) to get the n best results from the system. In our case

n=5. The joint source channel model considers the previous context of both the source and the target TUs. Based on equation (2) the model can be visualized as:

$$P(E|B) = \prod_{k=1}^{k} P(< e, b >_i |e_{i-1}) \quad (3)$$

i.e the joint source channel model tries to maximize the probability of the current English-Bengali TU pairs given the previous English-Bengali TU pair.

Apart from the joint source channel model, we have also trained and evaluated our system using the trigram model. The trigram model considers the context of both next and previous TUs of the source language text. The trigram model can be represented as:

$$P(E|B) = \prod_{k=1}^{k} P(< e, b >_i |e_{i-1}, e_{i+1}) \quad (4)$$

However, our evaluation results shows that the joint source channel model performs better transliteration than the trigram model.

6 Evaluation

We evaluate our proposed joint source channel model for English to Bengali machine transliteration system using the evaluation matrices defined in (Zhang et al., 2012). Based on the works of (Zhang et al., 2012), we have used the following four evaluation metrics[3]: a) Word-Accuracy in Top 1 (ACC) b) Fuzziness in top-1 (Mean F-Score) c) Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP_{ref}). We will define all these matrices in the following subsection. However, before that we define the following notations used in all the above evaluation matrices: N : Total number of words in the test sets; n_i : Number of reference transliterations for i-th word in the test set ($n_i >= 1$); $r_{i,j}$: j-th reference transliteration for i-th word in the test set; $c_{i,k}$: k-th candidate transliteration for i-th name in the test set ($1<k<10$); K_i : Number of candidate transliteration produced by a transliteration system

6.1 Word Accuracy in Top-1 (ACC)

The ACC or word error rate measures the correctness of the first transliteration output in the top 10 list generated by the transliteration system. ACC=1 implies that all top candidates are correct transliterations i.e. they match one of the references, and ACC = 0 means that none of the top candidates are correct. We mathematically represent ACC as:

$$ACC = \frac{1}{N} \sum_{i=1}^{N} \begin{Bmatrix} 1 \; if \; \exists r_{i,j} : r_{i,j} = c_{i,k} \\ or \\ 0 \; otherwise \end{Bmatrix}$$

[3] All these metrics are already defined in (Zhang et al., 2012). However, for the sake of readability, we are discussing these metrics once again in the paper.

6.2 Fuzziness in Top-1 (Mean F-Score)

The mean F-Score computes on an average, how different the top-1 transliterated output is from its closest reference. F-Score for each source word is a function of its precession and recall and equals 1 when the top candidate matches one of the references and 0 when there are no common characters between the candidate and any of the references.

The Precision and Recall values are calculated based on the length of the Longest Common Subsequence (LCS) between a candidate and a reference:

$$LCS(c,r) = \frac{1}{2}(|c| + |r| - ED(c,r))$$

Where, ED is the edit distance and |x| is the length of x (Zhang et al., 2012). Consequently, the recall, precision and F-score for i-th word are calculated as

$$R_i = \frac{LCS(c_{i,1}, r_{i,m})}{|r_{i,m}|};\ P_i = \frac{LCS(c_{i,1}, r_{i,m})}{|c_{i,m}|};\ F_i = 2 \times \frac{R_i \times P_i}{R_i + P_i}$$

Where, $r_{i,m}$ is the reference for which the edit distance has the minimum value. $r_{i,m}$ is defined as:

$$r_{i,m} = \arg\min_j \left(ED(c_{i,1}, r_{i,j})\right)$$

Here, the length is computed in terms of distinct Unicode characters. Note that unlike English or other European languages, Bangla has got severe complexities in terms of formations of characters. One such complexity lies in the formation of conjugate characters where the halant marker (্) is used to combine two or more characters to form a new one. For example, in the case of ক্ষ = ক + ্+ ষ. In our present work we have considered the halant marker (্) as a distinct character.

6.3 Mean Reciprocal Rank (MRR)

The mean reciprocal rank signifies the mean rank for any correct answer produced by the system, from the candidate set. The reciprocal rank or $1/MRR$ signifies approximately the average rank of the correct transliteration. MRR value close to 1 means that the correct answer is mostly produced close to the top of the n-best lists. The reciprocal rank (or RR) is defined as:

$$RR_i = \left\{ \begin{matrix} min_j \frac{1}{j}\ if\ \exists r_{i,j}, c_{i,k} : r_{i,j} = c_{i,k} \\ or \\ 0\ otherwise \end{matrix} \right\}$$

Consequently, the MRR can be mathematically expressed as:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N} RR_i$$

6.4 Mean Average Precision (MAP_{ref})

This measures the precision in the Top-n candidates for the i-th test word, for which reference transliterations are available. If all of the references are produced, then the MAP is 1. Considering the number of correct candidates for the i[th] test word in k-best list is $num(i,k)$. Therefore, MAPref can be expressed as:

$$\mathrm{MAP}_{\mathrm{ref}} = \frac{1}{N}\sum_{i}^{N}\frac{1}{n_i}\left(\sum_{k=1}^{n_i} num(i,k)\right)$$

7 Results

Both the proposed joint source channel model as well as the trigram models has been tested using a test corpus of 2000 words. The test corpus has been prepared manually by the experts. Further, each element of the test set is checked whether they are present in the training corpus. All the 2000 test words were presented to the models. The output generated by each model is further checked manually by the expert users. The performance of both the models has been evaluated based on the four different metrics discussed in the previous section. Table 1 reports the comparison between the performances of each of the models based on the individual metrics.

Table 1. Results of English-Bengali Translation Task Output

Model	ACC	F-Score	MRR	MAP_{ref}
Joint Source Channel Model	0.1221	0.2468	0.1833	0.139
Trigram Model	0.1293	0.7687	0.1839	0.287

Based on the reported data of table 1, we observed that the performance of the joint source channel model performs better that of the trigram model. We further observed that although the ACC and MRR for both the models are comparable (0.122 and 0.183for joint source channel model and 0.119 and .177 for tri-gram model respectively), the F-Score and MAP_{ref} score of the joint source channel model far exceeds that of the tri-gram model. Overall the joint source channel model surpasses the performance of the trigram model.

8 Conclusion

Transliteration is an extremely ambiguous process. The problem of back-transliteration is considered to be even more difficult than the forward transliteration. This is mainly due to the large number of ambiguities that are needed to be resolved

between the source and the target language words. In this paper we present the joint source channel model and a trigram based model to automatically transliterate Romanized Bengali texts to their respective Bengali form. Accordingly, we have tried to automatically extract the transliteration units (TU) and used the models for the training purpose. Finally, both the models have been tested using four different evaluation matrices. We have observed the performance of the joint source channel model surpasses the performance of the trigram model.

References

1. Ekbal, A., Naskar, S., Bandyopadhyay, S.: A modified joint source-channel model for transliteration. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp. 191–198 (2006)
2. Zhang, M., Li, H., Kumaran, A., Liu, M.: Report of NEWS 2012 machine transliteration shared task. In: Proceedings of the 4th Named Entity Workshop, pp. 10–20 (2012)
3. Schwartz, R., Chow, Y.L.: The N-best algorithm: An efficient and Exact procedure for finding the N most likely sentence hypothesi. In: Proceedings of ICASSP 1990, Albuquerque, pp. 81–84 (1990)
4. Knight, K., Graehl, J.: Machine Transliteration. Computational Linguistics 24(4), 599–612 (1998)
5. Ehara, Y., Tanaka-Ishii, K.: Multilingual text entry using automatic language detection. In: Proceedings of IJCNLP, pp. 441–448 (2008)
6. Lee, C.-J., Chang, J.S.: Acquisition of English-Chinese Transliteration Word Pairs from Parallel-Aligned Texts using a Statistical Machine (2003)
7. Translation Model. In: Proceedings of HLT-NAACL Workshop: Building and Using parallel Texts Data Driven Machine Translation and Beyond, Edmonton, pp. 96–103 (2003)
8. Li, H., Kumaran, A., Zhang, M., Pervouchine, V.: Report of NEWS 2009 Machine Transliteration Shared Task. In: Proceedings of ACL-IJCNLP 2009 Named Entities Workshop, pp. 1–18 (2009)
9. Goto, I., Kato, N., Uratani, N., Ehara, T.: Transliteration considering Context Information based on the Maximum Entropy Method. In: Proceeding of the MT-Summit IX, New Orleans, USA, pp. 125–132 (2003)
10. Li, H., Min, Z., Jian, S.: A Joint Source-Channel Model for Machine Transliteration. In: Proceedings of the 42nd Annual Meeting of the ACL (ACL 2004), Barcelona, Spain, pp. 159–166 (2004)
11. Lee, J.S., Choi, S.: English to Korean statistical transliteration for information retrieval. Computer Processing of Oriental Languages 12(1), 17–37 (1998)

Named Entity Recognition for Gujarati: A CRF Based Approach

Vipul Garg, Nikit Saraf, and Prasenjit Majumder

Information Retrieval Lab,
Dhirubhai Ambani Institute of Information and Communication Technology,
Gandhinagar, 382007
{201001049,201101218,p_majumder}@daiict.ac.in

Abstract. This paper is about Named Entity Recognition (NER) for Gujarati language. Not much work has been done in NER for Gujarati. In this paper, an NER tagger is build using Conditional Random Fields (CRF). The NER tagger is capable of identifying person, location and organization names with an F1-score of 0.832.

Keywords: NER for Gujarati, CRF, Named Entity, POS Tagger.

1 Introduction

Named Entities are one of the most important textual unit in the Information Extraction domain as they express an important part of the meaning of a document. NER is also an important part of Natural Language Processing (NLP) applications like Machine Translation and Text Summarization and is also used in search engines. For Gujarati language no such NER tagger exist and hence this paper propose a CRF based NER tagger using CRFsuite (Okazaki, 2007) to solve the problem[1].

The paper is organized as follows. Section 2 gives an overview of approaches to solve the NER problem, Section 3 talks about NER for Indian languages. Furthermore section 4 describes the Parts of Speech (POS) tagger, section 5 describes the NER Tagger based on CRF and finally section 6 concludes the paper.

2 Approaches to NER

There are various approaches to NER like (1) Handcrafted or automatically generated rules or patterns (2) look up from large lists or other resources (3) Statistical modeling of the language. In the first approach patterns and rules are used to find named entities by looking for common patterns which are present in a particular type of entities, for example company names in US generally ends with Inc. and, in English, all the entity names necessarily starts with a capital letter which reduces the possible candidates of named entities significantly and makes the recognition much easier. The

[1] This Project is supported by CLIA.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 761–768, 2013.
© Springer International Publishing Switzerland 2013

second approach uses lists of names of person, location and organization to check if a given word in the document is an entity. This method restricts the recognition of named entities to only those which are present in the lists. This require a good resource of named entities so that the tagger can tag most of the entities present in a document. These lists are also called gazetteers. This approach is fast and simple but its performance highly depends on the quality of gazetteers used and how often they are updated because named entities are large in number and they constantly evolve. Also this method does not solves ambiguities like whether to classify "Oxford" as a city or a University, which depends on the context in which it is used. This problem is well tackled by the third approach which statistically model the language patterns by analyzing when and where entity occurs in a document. Machine learning techniques are used for statistical modeling which can be either unsupervised, semi-supervised or supervised mode of learning. Unsupervised and semi-supervised mode of learning are used when there is a scarcity of annotated data for training but the best performance is obtained by using supervised mode of learning which requires a large amount of good quality annotated corpus.

3 NER for Indian Languages

Due to the advent of Machine learning techniques and algorithms, NLP research has taken giant steps all over the world and the accuracy of various NLP system has significantly improved. Majority of research work is done for English language which has a very high NER accuracy compared to Indian Languages. NLP tools for Indian languages have not yet reached that level because of lack of annotated data and lexical resources. Recorded accuracy for NER in Indian languages are shown in table 1. (Asif Ekbal R. H., 2008) (B. Sasidhar, 2011). Majority of Indian languages are highly inflected languages due to which many of the techniques used for English does not work here. For example lack of capitalization feature in Indian languages reduces a significant feature which is used by English in NER systems, as in English the Named entities always starts with a capital letter. The difficulty is further increased because of large overlap in proper noun and common nouns in Indian languages. For e.g. Anand, Vijay, Kiran etc. are common nouns as well as proper names. Large gazetteers are also not available for many Indian languages. Non availability of labeled data in Indian languages becomes a hurdle in supervised classification. Indian languages being highly inflected, provides rich and challenging sets of linguistic and statistical features resulting in long and complex word forms. Indian languages also have relatively free word order, and because of all these reasons they require language dependent feature sets.

Table 1. F1 score comparision for various Indian Languages

Language	Method used	F1-score (%)
Tamil	CRF	83.8
Bengali	CRF	90.7
Telugu	CRF	92.0
Gujarati	CRF	83.3

4 CRF Based POS Tagger

Parts of speech tagger is an NLP tool which tags the words in a document with their POS tag say noun, verb, adjective etc. POS tags can improve the results of an NER tagger and hence this section describes a POS tagger developed for Gujarati Language. There are several ways to develop a POS tagger, here a CRF based supervised learning method is used to for its development.

4.1 Corpus

A corpus with 25 Gujarati documents each with 1000 POS tagged sentences is used for training a CRF based POS tagger. This corpus was collected from TDIL India website.

The corpus is a manually tagged corpus developed under the ILCI Consortium. (Jha, 2010).

4.2 Feature Set

A feature set of 18 values is used as a template for the CRF based POS tagger which consist of Prefix of the word, Suffix of the word, length of word, Word context with its 15 values of various combinations as listed below in table 2. The accuracy of the POS tagger developed is 92.7%.

Table 2. Feature vector used in POS tagger

Word Suffix of length 3
Word Prefix of Length 3
Length
w[-2]
w[-1]
w[0]
w[1]
w[2]
w[-2]/w[-1]
w[-1]/w[0]
w[0]/w[1]
w[1]/w[2]
w[-2]/w[-1]/w[0]
w[-1]/w[0]/w[1]
w[0]/w[1]/w[2]
w[-2]/w[-1]/w[0]/w[1]
w[-1]/w[0]/w[1]/w[2]
w[-2]/w[-1]/w[0]/w[1]/w[2]

5 CRF Based NER for Gujarati

5.1 Corpus

A corpus of 225 articles from "Gujarat Samachar" newspaper manually tagged with PERSON, LOCATION and ORGANIZATION tags, is used for training and testing the CRF. In all the corpus contains 122201 words. Each article is preprocessed according to the requirement of CRF suite which needs a file in which each line has a single word and its NER tag separated with a white space, A new line represents end of a sentence and two consecutive new lines represents end of the corpus. Two processed files were created, one with BIO tags which shows multiword entities and another without it.

5.2 Evaluation Metrics

Two standard measures, Precision (P) and Recall (R) are used for evaluation of the Named Entity (NE) tagger, where precision is the measure of the number of entities correctly identified over the number of entities identified and recall is the measure of number of entities correctly identified over actual number of entities. F measure is calculated which is the harmonic mean of precision and recall

$$F = \frac{(\beta^2+1)PR}{\beta^2 R+P} \quad (1)$$

When $\beta = 1$, F measure is called F1 measure or simply F1 score.

F1 scores are calculated by performing 10-fold cross-validation which partition the training data in 10 complementary parts and then perform the analysis on one subset (called the training set), and validates the analysis on the other subset (called the validation set or testing set). The final estimation is the average of all ten rounds. Cross validation helps in assessing how the results of a statistical analysis will generalize to an independent data set.

5.3 Feature Set

Following are the features used while training.

- **Length:** It is observed that very short words are rarely noun so this feature is 1 if the word length is greater than or equal to three otherwise it is 0.
- **Affixes:** Prefix and suffix of words are used as features because they are observed to be helpful in recognition of entities in highly inflected languages like Gujarati. These affixes are not linguistically meaningful morpheme, Suffixes of length from 4 characters down to 1 character and prefixes of length from 7 characters down to 1 character are considered for each word, words with length smaller than three are given a value 'NA' as their affixes.

- **Position:** A binary feature which is one if the word is at the starting of the sentence otherwise it is zero. This feature is useful because most of the time the starting word of a sentence is a named entity.
- **Word Context:** Context of the word of window size four is used which takes two words before and two words after the word as feature. This helps modeling the language structure about how where and with which words entities are used in a sentence. There are total seven feature values for word context which includes the word itself, two words before it, and two words after it, and pairing of word with its previous and next word.
- **POS tag:** Parts of Speech (POS) tag of a word is also considered as a feature because all the entities are nouns, these tags are calculated using the POS tagger described in section 5.

A feature vector is generated for each word in the corpus containing features described above. Using these features of words and the output class (tag), CRF learns and generate a model which is later used in testing. The feature template used in the first two experiments is given in table 3

Table 3. Feature template used for training CRF based NER Tagger

w_{i-2}
w_{i-1}
w_i
w_{i+1}
w_{i+2}
Combination of $w_{i-1,}$ w_i
Combination of $w_{i,}$ w_{i+1}
Length
Prefix
Suffix
Position

5.4 Experiments and Results

Experiment 1

CRF is given training data tagged with four tags, PERSON, LOCATION, ORGANIZATION and OTHERS. This does not takes into account the multiword entities and each word is tagged as discrete entity. Feature template shown in table 3 is used.

Table 4 shows the results obtained when 10 fold cross-validation is performed using the feature set given in table 4 for the training data. The notations used in Table-3 are as follows: PER – person, LOC – location, ORG – organization, OTH – others and AVG – Macro average of all four tags.

Table 4. Results of Experiment-1

	Precision	Recall	F1-score
PER	0.88	0.70	0.78
LOC	0.86	0.75	0.80
ORG	0.84	0.67	0.75
OTH	0.98	0.99	0.98
AVG	0.89	0.78	**0.829**

So this NER tagger gives a F1-score of 0.829 as seen in table 4

Experiment-2

The second experiment extends the tagger to the recognition of multi word entities by using BIO Tags (Beginning, Intermediate and Others). For this the training data used contains 7 tags: B-PERSON, I-PERSON, B-LOCATION, I-LOCATION, B-ORGANIZATION, I-ORGANIZATION and OTHERS.

Results obtained after 10 fold cross-validation are given in table 5:

Table 5. Results of Experiment 2

	Precision	Recall	F1-score
B-PER	0.84	0.65	0.73
I-PER	0.69	0.60	0.64
B-LOC	0.86	0.76	0.81
I-LOC	0.79	0.51	0.62
B-ORG	0.87	0.63	0.73
I-ORG	0.86	0.63	0.72
OTH	0.97	0.99	0.98
AVG	0.84	0.68	**0.75**

The F1-score is significantly reduced here compared to experiment one because of increased number of tags for identifying multi word named entities. The lower score is mainly due to poor recall value of the tagger.

Experiment-3

The third experiment includes the POS tags of the word as a feature in the feature vector in addition to the features in table 3, The results for four tags, PERSON, LOCATION, ORGANIZATION and OTHERS and BIO tags are given in table 6 and table 7 respectively.

Table 6. Results of Experiment 3 with simplified tags

	Precision	Recall	F1-score
PER	0.87	0.71	0.78
LOC	0.87	0.76	0.81
ORG	0.86	0.67	0.76
OTH	0.98	0.99	0.98
AVG	0.89	0.78	**0.832**

Table 7. Results of Experiment with BIO tags

	Precision	Recall	F1-score
B-PER	0.85	0.66	0.74
I-PER	0.71	0.63	0.67
B-LOC	0.86	0.77	0.81
I-LOC	0.82	0.53	0.64
B-ORG	0.83	0.64	0.75
I-ORG	0.86	0.64	0.73
OTH	0.98	0.99	0.98
AVG	0.85	0.70	**0.761**

The F1-scores are improved by the usage of POS tags as one of the feature.

6 Conclusion

CRF models work very well for the highly inflective Indian languages, in fact better than other systems like HMM, MEMM etc. (Vijay Sundar Ram R, 2011). In the first attempt of developing an NER tagger, the F1- scores are low compared to that of other languages like English. Use of POS tag as a feature shows clear improvement in the F1-scores of the tagger. There are several ways in which this tagger can be improved further like by including gazetteers of person, location and organization names. Improving the Post tagger so that it can identify nouns in a much accurate way. Achieving a high performing NER system for Gujarati requires more study and deeper understating of linguistic features. Various permutation and combination of feature sets can be used and tested for getting high recall value and eventually higher F1-scores.

References

Asif Ekbal, R.H.: Language Independent Named Entity Recognition in Indian Languages. In: IJCNLP, pp. 33–40 (2008)

Asif Ekbal, R.H.: Named Entity Recognition in Bengali: A Conditional Random Field Approach. In: IJCNLP, pp. 589–594 (2008)

Sasidhar, B., P.Y.: A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu. IJCSI International Journal of Computer Science (2011)

David Nadeau, S.S. (n.d.).: A survey of named Entity recognition and classification. National Research Council Canada/ New York University

Jha, G.N.: LREC, The TDIL Program and the Indian Language Corpora Initiative(ILCI). ILCI (2010), retrieved from `http://www.lrec-conf.org/proceedings/lrec2010/summaries/874.html`

Okazaki, N.: CRFsuite: A fast implementation of Conditional Random Fields, CRFs (2007), retrieved from `http://www.chokkan.org/software/crfsuite/`

Singh, A.K.: Named Entity Recognition for South and South East Asian Languages: Taking Stock. Asian Federation of Natural Language Processing, Hyderabad (2008), retrieved from http://www.aclweb.org/anthology/I/I08/I08-0303

Vijay Sundar Ram R, P. R: Identification of Different Feature Sets for NER tagging using CRFs and its impact, vol. 11. Language in India, Chennai (2011)

How Word Order Affects Sentence Comprehension in Bangla: A Computational Approach to Simple Sentence

Manjira Sinha, Koustav Rudra, Tirthankar Dasgupta, and Anupam Basu

Indian Institute of Technology Kharagpur
{manjira,tirtha,anupam}@cse.iitkgp.ernet.in

Abstract. Sentence comprehension is an integral and important part of whole text comprehension. It involves complex cognitive actions, as a reader has to work through lexical, syntactic and semantic aspects in order to understand a sentence. One of the vital features of a sentence is word order or surface forms. Different languages have evolved different systems of word orders, which reflect the cognitive structure of the native users of that language. Therefore, word order affects the cognitive load exerted by a sentence as experienced by the reader. Computational modeling approach to quantify the effect of word order on difficulty of sentence understanding can provide a great advantage in study of text readability and its applications. Handful of works has done in English and other languages to address the issue. Bangla, which is the fifth mostly spoken languages in the world and a relatively free word order language, still does not have any computational model to quantify the reading difficulty of a sentence. In this paper, we have developed models to predict the comprehending difficulty of a simple sentence according to its different surface forms in Bangla. In the course of action, we have also established that difficulty measures for English do not hold in Bangla. Our model has been validated against an extensive user survey.

Keywords: sentence comprehension, word order, surface forms, Bangla, computational model, simple sentence.

1 Introduction

Complexity of a sentence is the amount of effort a user needs to put in order to understand or comprehend the sentence. Sentence complexity is an important factor in accessing text readability, language acquisition and language impairment. When a reader scans (generally left to right) a sentence, she first processes the syntax (structure and word organization) and semantics (meaning represented by the words) and then reduces them to a semantic whole to store in the memory (Levy, 2013). The short-term memory of the reader engages in real time comprehension of a sentence. While processing a sentence, the short-term memory encounters two types of costs (Oya, 2011): storage cost of the structure built in memory so far and the integration cost due to the proper insertion of the current word into that structure. Therefore, the integration complexity depends upon the relative positions of the entities to be connected, i.e., word order of the sentence.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 769–779, 2013.
© Springer International Publishing Switzerland 2013

As discussed above, one of the important grammatical information for sentence interpretation is the word order as it determines the organizations of different grammatical features. It has great impact on the sentence complexity (Meltzer et al., 2010) as it influences both the storage and integration cost and expectation load. Different languages follow different construction rules to build sentences and thus different word orders. Research has been performed to the study effect of word ordering in sentence comprehension in languages like English, Finnish, German (SWINNEY, 1998; Weyerts et al., 2002; Kaiser and Trueswell, 2004). In this paper, the language concerned is Bangla. Bangla is a descendant of the Eastern Indo-Aryan language family[1]. Typologically, it is an inflexional analytic language. Syntax or Sentence structure of Bangla differs from English in many aspects. Bangla is a head final language where the principle word order is subject-object-verb (SOV). It is also a relatively free word-order language as it permits free word order in its constituent chunk or local word group level. Intra-chunk reordering of words is not always permitted; different surface forms of the same sentence are possible, which are grammatically correct; some surface forms are easy to comprehend and some are difficult. Therefore, even simple sentences in Bangla (Chatterji, 1926) can have different surface forms with different comprehending complexities. Till date, there is no prominent study to computationally model the cognitive load associated with different word orders in Bangla.

In this study, our objective is to develop model to quantify the varying reading difficulties of different surface forms of the same sentence. We have considered simple sentences i.e. sentences having one finite verb[2]. Simple sentences in Bangla can contain many language specific constructs. We have explored the underlying factors responsible for the differences in complexity among different surface forms, such as relative order of subject(s) and object(s) with respect to the verb and organization of non-finite structures. First, we have conducted an empirical user survey, and then we have developed and enhanced our model to reflect the comprehending difficulty experienced by the readers efficiently. In the due course, we have demonstrated that although average dependency distance measure (ADD) (Oya, 2011) works well for English, it is not a good estimator of sentence difficulty in Bangla. Our proposed model takes into account both the relative position and number of unprocessed dependencies at an instant; it is unprocessed dependencies that give rise to expectation gaps in user's cognition. Thus, it models both storage and integration costs on reader's short-term memory in processing a sentence based on different surface forms. We have found high correlation among user preferences and model predictions.

The paper is organized as follows: section 2 states the related works, section 3 describes the user study, results and discussions, section 4 contains conclusion and future works.

2 Related Work

A handful of researches have been performed on sentence complexity and word order preference in sentence comprehension. Some approaches are based on dependencies such as placement of verbs in a sentence, position of subject and auxiliary verb in a

[1] http://en.wikipedia.org/wiki/Bengali_language
[2] http://en.wikipedia.org/wiki/Simple_sentence

sentence etc. Several psycholinguistic experiments have been performed to study the role of word order in sentence comprehension.

Dependency Locality Theory (DLT)(Gibson, 2000) states that there is a preference for closely related words to be close together in the sentence. Based on the theory, measures like Average Dependency Distance (ADD) (Oya, 2011) have been proposed to model the effect of distance between dependent words on sentence readability. The role of canonical word ordering (SWINNEY, 1998), effect of discourse contexts on preferred word ordering (Kaiser and Trueswell, 2004) and the role of syntactic and semantic cues on the reading load associated with different surface forms (Casado et al., 2005) have been studied by experts. Weyerts et al (Weyerts et al., 2002) examined the question of whether the human comprehension device exhibits word order preferences during on-line sentence comprehension. The focus was on the positioning of finite verbs and auxiliaries relative to subjects and objects in German. Results from the experiments (using self-paced reading and event-related brain potentials) showed that native speakers of German prefer to process finite verbs in second position, i.e. immediately after the subject and before the object. In Finnish, it have found that canonical word orders (S (subject) V (verb) O (object)) are felicitous in certain discourse contexts (subject – old + object – old), (subject - new + object - new), (subject - old + object - new), whereas, non-canonical ones are more convenient when object is old and subject is new (Kaiser and Trueswell, 2004). Results from neuro-imaging experiments showed that differences in canonical word order between SVO & SOV languages affect brain activation during sentence comprehension and caused a different load in the working memory process (Hashimoto et al.,). Recent studies are focusing on how different languages and different cognitive systems have adopted preference for different word orderings (Langus and Nespor, 2010; Gibson et al., 2013; Hall et al., 2013). Till now, no such prominent work is present in Bangla.

3 Computational Modeling of Reading Difficulty of Different Surface Forms

This section presents the computational approach to quantify the relative comprehension difficulty of different word-ordered forms of a Bangla sentence. It is divided in three subsections: first part briefly describes the user study, second part details the processing of the experimental sentences prior to model building, in third subsection, we have presented a detail stepwise account of model development and the last part analyses the results.

3.1 User Study

Sentence Selection

Seventy (70) test sentences (simple) were selected randomly from a corpora made of wide range of documents: Bangla blogs, short stories by eminent writers in this language, Bangla news articles and textbook contents. The sentence length was between 5-20 words. Three different surface forms were constructed corresponding to each sentence.

Participants and Procedure

Twenty-five (25) native speakers of Bangla of age group from 17-25 years participated in this study. Their educational background ranges from plus 2 level of school to undergraduate in college; they all have Bangla as their first language in their school syllabus.

The procedure is straightforward: each participant were presented with the three surface forms of the same 70 sentences and were asked to relatively rank the three forms corresponding to each sentence based on the comprehension difficulty experienced by them. The users were given a maximum of 30 seconds per sentence to express their decisions. The time constraint has been imposed to minimize user bias. We present some sentences in Table1.

Table 1. Example from user experiment (1 = easy, 2 = medium, 3= hard)

No	Form Id	Sentence[3]	User Rating
1	$\mathbb{S}_1$	মুখুজ্যের হাত থেকে পাথরের বাটি দুম করে পড়ে গেল mukhujyera hAta theke pAtharera bATi duma kare paDe gela (The stone bowl suddenly dropped from Mukhujje's hand)	1
	$\mathbb{S}_2$	মুখুজ্যের হাত থেকে পড়ে গেল দুম করে পাথরের বাটি mukhujyera hAta theke paDe gela duma kare pAtharera bATi	3
	$\mathbb{S}_3$	মুখুজ্যের হাত থেকে দুম করে পড়ে গেল পাথরের বাটি mukhujyera hAta theke duma kare paDe gela pAtharera bATi	2

3.2 Identification and Construction of Dependency Tree

To identify the different kinds of dependencies among the words of a sentence, we have used a hybrid dependency parser for Bangla(Dhar et al., 2012). It contains 26 Part-of-speech (POS) categories, 11 chunk types and 28 dependency tag set. We construct the dependency tree as follows:

Each word $\mathbb{w}$ in a sentence is a node in the tree. When a dependency relation relates from a node $\mathbb{w}_i$ to node $\mathbb{w}_j$, then we say node $\mathbb{w}_i$ is dependent on node $\mathbb{w}_j$. $\mathbb{w}_i$ and $\mathbb{w}_j$ are connected by a directed edge from $\mathbb{w}_j$ to $\mathbb{w}_i$ and $\mathbb{w}_j$ is considered at a higher level than $\mathbb{w}_i$. The node or word that is not dependent on other nodes is called the root or starting node, which is the main finite verb in our case. The root node is at level 1, subsequently, the nodes dependent on root node are at level 2. In this way, nodes that are dependent on nodes at level L are assigned level L+1. If any $\mathbb{w}_i$ from level L+2 is dependent on nodes in both L and L+1, then we only consider the dependency from L+2 to L+1.

[3] Sentences are written as Bangla-*iTrans transliteration* (Translation).

Example[4]**:** In the sentence **তোর সঙ্গে আমার একটা কথা আছে** – ***wora safge AmAra ekaTA kaWA Ace*** (I want to have a word with you), the *karta* (subject [k1]) is ***kaWA***, *karma* (object [k2]) is ***wora***, and the finite verb [root] is ***Ace***; ***AmAra*** is the possessive noun [r6], ***ekaTA*** is noun modifier [jnmod], ***safge*** is post-position of the object [ppl]. The dependency tree will look like:

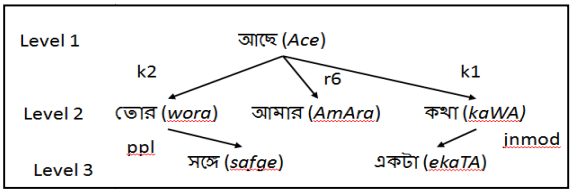


Fig. 1. Dependency tree of a sentence

3.3 Model Building

Testing Average Dependency Distance (ADD) Measure

ADD of a sentence is the sum of distance of all the dependencies in the sentence divide by the number of dependencies of the sentence. According to the dependency locality theory, the lower the distance between the dependencies in a sentence, the easier is to comprehend. At first, we have ADD to check whether it could explain the relative preferences of the user. However, as can be seen from table 2 below, ADD fails to reflect the desire objective in case of Bangla. The disagreement on the values can be attributed to the differences in the sentence structure of English and Bangla.

Table 2. Comparison of ADD and user rating

No.	Form Id	Sentence	User rating	ADD rating
1.	$\mathbb{S}_1$	মোটামুটি আমার ব্যাপারে তারা সুপরিকল্পিত নীরবতা অবলম্বন করেছেন– moTAmuTi AmAra byApAre tArA suparikalpita nIrabatA abalambana karechhena (They have almost adapted a planned silence in my matters)	**1**	**2.28**
	$\mathbb{S}_2$	মোটামুটি তারা সুপরিকল্পিত নীরবতা অবলম্বন করেছেন আমার ব্যাপারে- moTAmuTi tArA suparikalpita nIrabatA abalambana karechhena AmAra byApAre	**2**	**3**
	$\mathbb{S}_3$	মোটামুটি তারা অবলম্বন করেছেন সুপরিকল্পিত নীরবতা আমার ব্যাপারে- moTAmuTi tArA abalambana karechhena suparikalpita nIrabatA AmAra byApAre	**3**	**2**

[4] Sentence written as **Bangla-*WX format*** (Translation); type of dependency is shown in third bracket.

Model for Sentence Hardness in Bangla

As already discussed before, different word ordering affects the storage, integration and expectation load of processing a sentence. Our assumption is that at any given instance, the difficulty of processing a word w in a sentence depends on two factors: number of unprocessed nodes or words dependent on w: exerts a storage cost on short-time memory and distance of the unprocessed dependencies of w: exerts a integration cost on the cognitive faculty.

To incorporate above points, we have used a dependency tree based weighted cost approach. As shown in fig 1, we have assigned different levels to the nodes of the dependency tree of a sentence from top to bottom; now we have assigned a cost with each level starting from bottom to top i.e, $cost(L) = \max(L) - L$, where L is the level number of the current node and $\max(L)$ is the maximum level number, then cost(L) represents the cost associated with that level. All nodes ($\mathbb{w}$) in the same level are said to have the same cost [$cost(L_{\mathbb{w}})$].

Before processing a node at higher level (low-level value), if a reader processes its dependents at lower levels (high-level value), comprehending the sentence should become easy to the reader, as influence of the high-level nodes on sentence comprehension is greater than that of low level. Also, in our model, we have considered a group of closely related words, which moves as a unit across a sentence in different surface form as a unit, such as postpositions (attached to noun or pronoun), nonfinite verb (attached with the main verb) and polar and vector in compound verbs. The algorithm works as follows:

- **Step 1:** A input sentence $\mathbb{S}$ is scanned from left to right and its dependency tree is constructed. We employ a left to right search strategy on the dependency tree.
- **Step 2:** Level numbers and level cost are assigned according to the rule stated above.
- **Step 3:** Starting from level $L = 1$, the model checks if there remain any unprocessed dependencies $d_{\mathbb{w}_i}$ (at level $L_{\mathbb{w}} + 1$) of the word $\mathbb{w}$ in the present level.
  — $d_{\mathbb{w}_i}$ add it to the list $unprocessed(\mathbb{w})$
- **Step 4:** $\forall\ d_{\mathbb{w}_i} \epsilon\ unprocessed(\mathbb{w})$,
  — the positional difference of $d_{\mathbb{w}_i}$ from $\mathbb{w}$ $[dis(d_{\mathbb{w}_i})]$ is calculated;
  — $cost(d_{\mathbb{w}_i}) = [dis(d_{\mathbb{w}_i})] * cost(L_{\mathbb{w}})$
  — $cost(d_{\mathbb{w}_i})$ is the comprehension overhead on $\mathbb{w}$ due to $d_{\mathbb{w}_i}$
- **Step 5:** Repeat steps 3 and 4 for all L and all $\mathbb{w}$
- **Total comprehension overhead on** $\mathbb{w}$ due to dependency gap is $Load(\mathbb{w}) = \sum_i cost(d_{\mathbb{w}_i})$
- **Net comprehension load of** $\mathbb{S}$ **is** $\boldsymbol{Load}(\mathbb{S}) = \sum \boldsymbol{Load}(\mathbb{w}), \forall\ \mathbb{w}\ \boldsymbol{\epsilon}\ \mathbb{S}$

Results and Discussion

The working of the model is illustrated below with the first example sentence and its three surface forms (refer table 2):

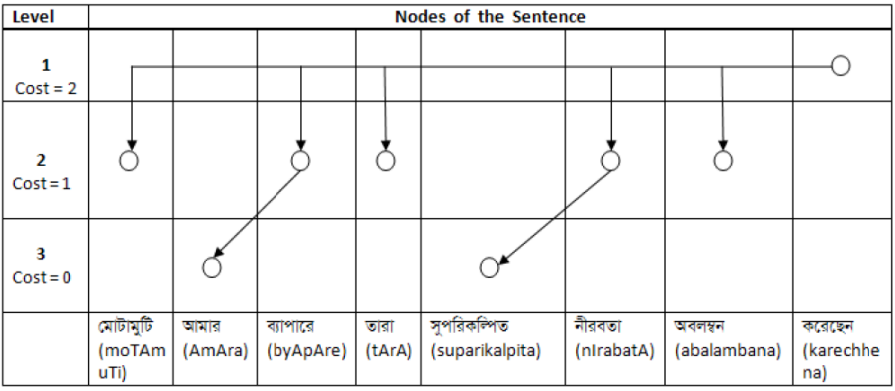


Fig. 2. Dependency tree for $\mathbb{S}_1$

Table 3. Cost estimation

Node	মোটামুটি moTAmuTi	আমার AmAra	ব্যাপারে byApAre	তারা tArA	সুপরিকল্পিত suparikalpita	নীরবতা nIrabatA	অবলম্বন abalambana	করেছেন karechhena
Cost	1	0	1	1	0	1	1	2
Unprocessed Nodes	Nil	Nil	Nil	Nil	Nil	Nil	Nil	Nil
Dependency Gap	0	0	0	0	0	0	0	0

Therefore, $Load(\mathbb{S}_1) = 0$

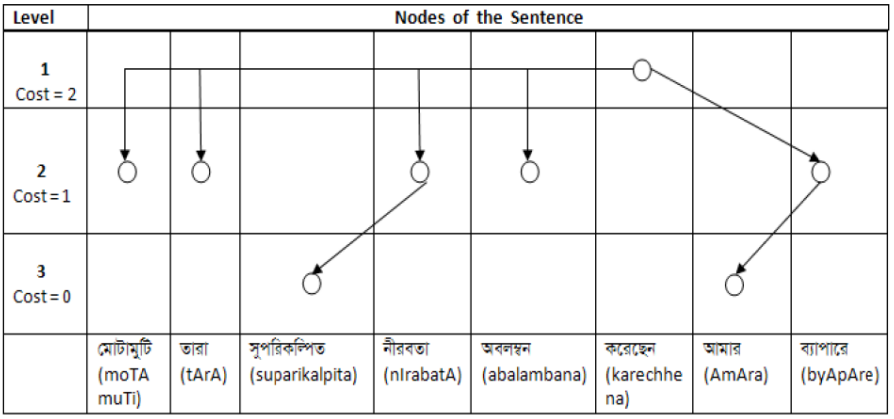


Fig. 3. Dependency tree for $\mathbb{S}_2$

Table 4. Cost Estimation (here আমার ব্যাপারে (AmAra byApAre) has been processed as an unit w.r.t. করেছেন (karechhena)

Node	মোটামুটি moTA muTi	তারা tAr A	সুপরিকল্পিত suparika lpita	নীরবতা nIrab atA	অবলম্বন abalam bana	করেছেন karech hena	আমার Am Ara	ব্যাপারে byAp Are
Cost	1	1	0	1	1	2	0	1
Unprocessed Nodes	Nil	Nil	Nil	Nil	Nil	আমার ব্যাপারে AmAra byApAr e	Nil	Nil
Dependency Gap	0	0	0	0	0	2*2=4	0	0

Therefore, $Load(\mathbb{S}_2) = 4$

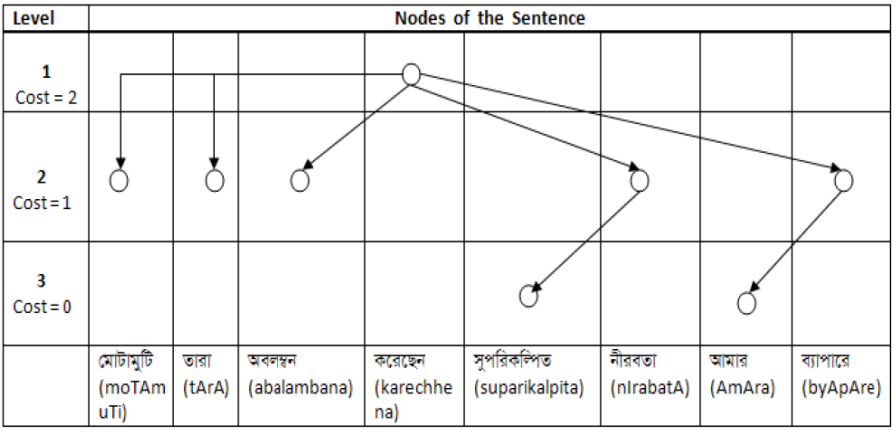


Fig. 4. Dependency tree for $\mathbb{S}_3$

Table 5. Cost estimation

Node	মোটামুটি moTAm uTi	তারা tAr A	অবলম্বন abalamb ana	করেছেন karechhena		সুপরিকল্পিত suparikal pita	নীরবতা nIrab atA	আমার AmA ra	ব্যাপারে byAp Are
Cost	1	1	1	2		0	1	0	1
Unprocessed Nodes	Nil	Nil	Nil	নীরবতা nIraba tA	আমার ব্যাপারে AmAra byApA re	Nil	Nil	Nil	Nil
Dependency Gap	0	0	0	2*2=4	4*2= 8	0	0	0	0

Therefore, $Load(\mathbb{S}_3) = 4 + 8 = 12$

Table 6. Comparison of rating predicted by the proposed model and the native readers.

No	Form Id	Sentence	User Rating	Proposed Model Rating
1	$\mathbb{S}_1$	মুখুজ্যের হাত থেকে পাথরের বাটি দুম করে পড়ে গেল mukhujyera hAta theke pAtharera bATi duma kare paDe gela (The stone bowl suddenly dropped from Mukhujje's hand)	**1**	**0**
	$\mathbb{S}_2$	মুখুজ্যের হাত থেকে পড়ে গেল দুম করে পাথরের বাটি mukhujyera hAta theke paDe gela duma kare pAtharera bATi	**3**	**10**
	$\mathbb{S}_3$	মুখুজ্যের হাত থেকে দুম করে পড়ে গেল পাথরের বাটি mukhujyera hAta theke duma kare paDe gela pAtharera bATi	**2**	**4**
2	$\mathbb{S}_1$	প্রথম সন্তান মেয়ে হওয়ায় আদিত্যর ইরিটেশন প্রায় মাপার বাইরে চলে যাচ্ছিল prathama santAna meYe haoYAYa Adityara iriTeshana prAYa mApAra bAire chale yAchchhila (Aditya was very irritated on his first child being a daughter)	**1**	**0**
	$\mathbb{S}_2$	প্রথম সন্তান মেয়ে হওয়ায় আদিত্যর ইরিটেশন চলে যাচ্ছিল প্রায় মাপার বাইরে prathama santAna meYe haoYAYa Adityara iriTeshana chale yAchchhila prAYa mApAra bAire	**2**	**6**
	$\mathbb{S}_3$	মাপার প্রায় বাইরে চলে যাচ্ছিল আদিত্যর ইরিটেশন প্রথম সন্তান মেয়ে হওয়ায় mApAra prAYa bAire chale yAchchhila Adityara iriTeshana prathama santAna meYe haoYAYa	**3**	**10**
2	$\mathbb{S}_1$	চার বছর ধরে তাড়া করে চলেছে কথাটা আমায় উন্মাদের মত chAra bachhara dhare tA.DA kare chalechhe kathATA AmAYa unmAdera mata	**3**	**14**
	$\mathbb{S}_2$	চার বছর ধরে কথাটা আমায় উন্মাদের মত তাড়া করে চলেছে chAra bachhara dhare kathATA AmAYa unmAdera mata tA.DA kare chalechhe	**1**	**0**
	$\mathbb{S}_3$	চার বছর ধরে উন্মাদের মত তাড়া করে চলেছে কথাটা আমায় chAra bachhara dhare unmAdera mata tA.DA kare chalechhe kathATA AmAYa	**2**	**6**

As discussed in the introduction section, our objective in this work is to develop a suitable computational model for relative hardness associated with different word ordering of a sentence. Now, comparing table 2 with table 3, table 4 and table 5, we can see that the approach adopted in this study clearly reflect the variation of comprehension difficulty experienced by the user according to different surface forms of the same sentence i.e., $Load(\mathbb{S}_1) < Load(\mathbb{S}_2) < Load(\mathbb{S}_3)$. As can be observed from the user preferences as well as from the model, while processing a word $\mathbb{w}$ of a sentence, if all dependent words of $\mathbb{w}$ are already present in our short-term memory then processing of the sentence (as in $\mathbb{S}_1$) becomes easy compared to the case (as in $\mathbb{S}_2$ and $\mathbb{S}_3$) when they are not present. While processing a sentence, readers break it into chunks and load it into her memory; chunking becomes easy if the above-mentioned condition holds. Table 6 below represents the difficulty ratings predicted by our model as well as by native reader for the sentences in table 1 and table 2. According to this model if a higher-level node (low-level value) is processed before processing its dependent nodes, which are at lower level (high-level value), then it incurs high reading cost on the reader. User study also demonstrates the same thing.

For almost every test case we took, our proposed model correctly predicts the cognitive load exerted on the user by change in word ordering of a sentence. However, we observe that sometimes users do not distinguish between two orderings. This may be due to the familiarity of the particular construct. Consider the following two constructs

1. (a) মুখের পানে চাহিয়া [mukhera pAne chAhiYA] and (b) চাহিয়া মুখের পানে [chAhiYA mukhera pAne]
2. (a) ভেতর থেকে বেরিয়ে [bhetara theke beriYe] and (b) বেরিয়ে ভেতর থেকে [beriYe bhetara theke]

In the above, two constructs 1(a) and 2(a) appear as easy to users than 1(b) and 2(b). As already stated, Bengali is relatively free order language many surface forms are possible and we have to check user preferences over those orderings and incorporate it into our model.

4 Conclusion and Future Works

In this paper, we have proposed a computational model to predict reading or comprehension experienced by different surface forms of a sentence in Bangla. Results from table 2 and table 6 demonstrate the fact that native readers of Bangla generally prefer to process finite verb at the end of a sentence. Also, within a nonfinite structure users find easy to process nonfinite verbs at the end of the structure. We have also observed from the experiment that as nonfinite structures are dependent on main finite verb users feel comfortable to process nonfinite structures before processing the finite verb. For example, দুম করে পড়ে গেল [duma kare pa.De gela] is easy to process than পড়ে গেল দুম করে [pa.De gela duma kare]. We have established that in case of Bangla, average dependency distance fails to distinguish among different surface forms of a sentence based on their hardness; we have also cited some potential explanations for the phenomena. According to the best of the knowledge of the authors, this is the first work on Bangla attempting to model computationally the cognitive load associated with sentence comprehension.

Up to this point, in our model, we have considered the distance and number of unprocessed dependencies of every word 𝕨 in a sentence in a level-wise approach. We have treated all types of dependencies as equal and not emphasized their relative misplacements in a surface form. In future, the model can be extended to a weighted dependency relation approach based on the amount of influence a particular type of dependency exerts in a sentence and the comprehension hardness arising from its shift in position from the ideal place. We also have to incorporate more parameters such as familiarity to a particular ordering in a context to better model the reading difficulty. Moreover, here we have considered the simple sentences. In future, we would like to take into account all three types of sentence (simple, complex and compound) and compare the specificities in the required modeling approaches.

References

1. Casado, P., MartÃ-n-Loeches, M., MuÃ±oz, F., FernÃ!'ndez-FrÃ-as, C.: Are semantic and syntactic cues inducing the same processes in the identification of word order? Cognitive Brain Research 24(3), 526–543 (2005)
2. Chatterji, S.-K.: The origin and development of the Bengali language, vol. 2. Calcutta University Press (1926)
3. Dhar, A., Chatterji, S., Sarkar, S., Basu, A.: A hybrid dependency parser for bangla. In: 24th International Conference on Computational Linguistics, p. 55 (2012)
4. Gibson, E.: The dependency locality theory: A distance-based theory of linguistic complexity. Image, Language, Brain, 95–126 (2000)
5. Gibson, E., Piantadosi, S.T., Brink, K., Bergen, L., Lim, E., Saxe, R.: A noisy-channel account of crosslinguistic word-order variation. Psychological Science (2013)
6. Hall, M.L., Mayberry, R.I., Ferreira, V.S.: Cognitive constraints on constituent order: Evidence from elicited pantomime. Cognition 129(1), 1–17 (2013)
7. Hashimoto, Y., Yokoyama, S., Kawashima, R.: Neuro-typology of sentence comprehension: Cross-linguistic difference in canonical word order affects brain responses during sentence comprehension. The Open Medical Imaging Journal
8. Kaiser, E., Trueswell, J.C.: The role of discourse context in the processing of a flexible word-order language. Cognition 94(2), 113–147 (2004)
9. Langus, A., Nespor, M.: Cognitive systems struggling for word order. Cognitive Psychology 60(4), 291–318 (2010)
10. Levy, R.: Memory and surprisal in human sentence comprehension (2013)
11. Meltzer, J.A., McArdle, J.J., Schafer, R.J., Braun, A.R.: Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. Cerebral Cortex 20(8), 1853–1864 (2010)
12. Oya, M.: Syntactic dependency distance as sentence complexity measure. In: Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics, pp. 313–316 (2011)
13. Swinney, D.A.: The influence of canonical word order on structural processing. Syntax and Semantics 31, 153–166 (1998)
14. Weyerts, H., Penke, M., Münte, T.F., Heinze, H.-J., Clahsen, H.: Word order in sentence processing: An experimental study of verb placement in german. Journal of Psycholinguistic Research 31(3), 211–268 (2002)

Importance of Utterance Partitioning in SVM Classifier with GMM Supervectors for Text-Independent Speaker Verification

Nirmalya Sen[1], Hemant. A. Patil[3], Shyamal Kr. Das Mandal[1], and K. Sreenivasa Rao[2]

[1] Signal Processing Group, C.E.T, IIT Kharagpur, India
[2] S.I.T, IIT Kharagpur, India
[3] DA-IICT, Gandhinagar, Gujarat, India
nirmalya75@rediffmail.com, hemant_patil@daiict.ac.in,
sdasmandal@cet.iitkgp.ernet.in, ksrao@sit.iitkgp.ernet.in

Abstract. This paper compares performances between GMM-UBM classifier and SVM classifier with GMM supervector as the linear kernel for text-independent speaker verification. The MFCC feature set has been used for this comparison. Experimental evaluation was conducted on the POLYCOST database. The importance of utterance partitioning for training speech has been discussed. Results reveal that, without utterance partitioning, the accuracy of SVM classifier with GMM supervectors for small test segment is poor. For proper utterance partitioning of the training speech, the SVM classifier with GMM supervectors performs significantly better compared to GMM-UBM baseline. The detailed derivation of GMM supervector has also been discussed.

Keywords: Speaker verification, GMM-UBM, GMM supervectors, Utterance partitioning.

1 Introduction

For speaker verification, the classical approach is Gaussian Mixture Model-Universal Background Model (GMM-UBM). Here we first prepare a speaker independent background model called UBM. After that from the training speech of the speaker, using maximum a posteriori (MAP) adaptation, we derive the speaker models. Though the GMMs are very simple, they provide effective likelihood functions. Then using the log-likelihood ratio between the speaker model and the background model, we calculate the score of test utterance [1].

To explore the knowledge of discriminative training in the area of speaker verification, people have tried support vector machine (SVM) [2]. However it is known that, there are some training problems of SVM for large data set. Particularly for large data set the training time of SVM is very long. Therefore the meaningful reduction of training set is very essential. Initially people tried vector quantization based approach to reduce the training data. Using vector quantization, they calculated the means of various clusters and used them to train the SVM. Later it was observed that, SVM

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 780–789, 2013.
© Springer International Publishing Switzerland 2013

with GMM supervector as the linear kernel [3], performed quite better compared to the previous approach. Presently SVM with GMM supervector is the most widely used classifier for most of the speaker verification system.

However, in supervector-based classifier, there is a mismatch between training supervector and test supervector. This is due to the fact that MAP adaptation depends on the length of speech signal .This mismatch is very prominent for short duration test speech. This point did not receive much attention on the NIST speaker verification experiment, because the test utterances of NIST database are not very small.

It is possible to effectively handle the above problem by utterance partitioning of the training speech [4]. Here instead of adapting the UBM with complete speech, we first divide the speech signal into sub-utterances. From each sub-utterance, using MAP adaptation of UBM, we generate one supervector. As the length of training utterance has been reduced, the mismatch between training supervector and test supervector will also reduced for small test speech.

The paper is organized as follows. Section 2 discusses about classical GMM-UBM speaker verification system. Section 3 introduces the concept of support vector machine (SVM). In Section 4, we have given the detail derivation of construction of GMM supervector. The concept of utterance partitioning has been introduced in Section 5. In Section 6, we report our experimental results based on the POLYCOST database [5]. Results show that, the utterance partitioning can reduce the equal error rate (EER) significantly for GMM supervector based classifier. Concluding remarks are given in Section 8.

2 Gaussian Mixture Model-Universal Background Model

In this section we briefly discuss the classical GMM-UBM system for speaker verification.

2.1 Universal Background Model

In the GMM-UBM system, a single speaker-independent background model is used. The UBM is a large GMM trained to represent the speaker-independent distribution of feature vectors [1]. Therefore, UBM should reflect the expected alternative speech to be encountered during verification. This applies to the type and quality of speech, as well as the composition of speakers. If it is known a priori that, male speakers will be tested only with male speech and female speakers will be tested only with female speech, then it is preferable to train one UBM for male speakers and separate UBM for female speakers. In general, where there is no prior knowledge of the gender composition of the alternative speakers, the UBM should be trained using gender-independent speech.

2.2 Adaptation of Speaker Model

The speaker models in the GMM-UBM system are generated by adapting the parameters of the UBM using the speaker's training speech with the help of the maximum a posteriori (MAP) estimation [1]. MAP adaptation is a two-step estimation process.

The first step is identical to the expectation step of the EM algorithm. Here the sufficient statistics of the speaker's training data are computed for each mixture of the UBM. In the second step of the MAP adaptation, these new sufficient statistics are then combined with the old sufficient statistics from the UBM mixture parameters using a data-dependent mixing coefficient. Hence the mixtures with high counts of training data from the speaker rely more on the new sufficient statistics for final parameter estimation and mixtures with low counts of training data from the speaker rely more on the old sufficient statistics for final parameter estimation. The concise discussion of parameter adaptation process is given below,

We started with a UBM which is a well-trained Gaussian mixture model. From the training speech of the hypothesized speaker, the sequence $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ of feature vectors is extracted. Then we determine the probabilistic alignment of these feature vectors into the UBM mixture components. Therefore for acoustic class i of UBM we calculate following a posteriori probability for all the feature vectors as given below,

$$\Pr(i \mid \mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j}^{M} w_j p_j(\mathbf{x}_t)} \tag{1}$$

Using the above a posteriori probability, we calculate sufficient statistics for mean, variance and weight for the acoustic class i as given below,

$$n_i = \sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t)\,,\ E_i(\mathbf{x}) = \frac{1}{n_i}\left(\sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t)\mathbf{x}_t\right) \quad \text{and} \quad E_i(\mathbf{x}^2) = \frac{1}{n_i}\sum_{t=1}^{T} \Pr(i \mid \mathbf{x}_t)(\mathrm{diag}((\mathbf{x}_t)^t \mathbf{x}_t)) \tag{2}$$

For each mixture, a data-dependent adaptation coefficient α_i is used. The definition of α_i is given below,

$$\alpha_i = \frac{n_i}{n_i + r} \tag{3}$$

Here r is a relevance factor which controls the degree of adaptation.

The above sufficient statistics from the training data are used to update the old sufficient statistics of the UBM for the acoustic class i as given below,

Let $\boldsymbol{\mu}_i$ be the mean vector of the acoustic class i of UBM. Then the adapted mean vector $\hat{\boldsymbol{\mu}}_i$ will be as follows,

$$\hat{\boldsymbol{\mu}}_i = \alpha_i E_i(\mathbf{x}) + (1 - \alpha_i)\boldsymbol{\mu}_i \tag{4}$$

2.3 Log-Likelihood Ratio Computation

At the time of testing, from the speech utterance of the hypothesized speaker, the sequence $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_T\}$ of feature vectors is extracted. The log-likelihood ratio score for the test utterance is calculated as given below,

$$\Lambda(\mathbf{Y}) = \frac{1}{T}(\log(p(\mathbf{Y} \mid \lambda_{spk})) - \log(p(\mathbf{Y} \mid \lambda_{UBM}))) \tag{5}$$

3 Support Vector Machine Classifier

An SVM is a two-class classifier based on a hyperplane separation concept [2]. It is constructed from sums of a kernel function $K(\cdot,\cdot)$ as given below,

$$f(\mathbf{x}) = \sum_{k=1}^{M} \lambda_k t_k K(\mathbf{x}, \mathbf{x}_k) + d \tag{6}$$

Here t_k are the ideal outputs belong to either $+1$ or -1. The coefficients λ_k are the Lagrange multipliers. The constraints are $\sum_{k=1}^{M} \lambda_k t_k = 0$, and $\lambda_k > 0$. The vectors $\mathbf{x}_k$ are the support vectors and obtained from the training data sets by an optimization process [6]. All the parameters of the above equation (i.e. the Lagrange multipliers and d) are obtained through a quadratic programming optimization problem [6].The optimization condition is based upon a maximum margin concept. A hyperplane is placed between the two training data sets in such a way that, the separation between the training data sets and the hyperplane is maximum. The training vectors which lie closest to the separator hyperplane are the support vectors of the above equation.

The kernel function $K(\cdot,\cdot)$ is designed in such a way that, it obeys the Mercer condition [2]. The kernel is required to be symmetric positive semi-definite. The Mercer condition ensures that the optimization problem of SVM is bounded, and the margin concept is valid. The kernel function $K(\cdot,\cdot)$ can be expressed in the following inner product form,

$$K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle \tag{7}$$

Here Φ is a mapping from the input vector space to a possible infinite dimensional vector space.

4 Construction of GMM Supervector

We started with a Gaussian mixture model universal background model (GMM-UBM),

$$p(\mathbf{x}) = \sum_{k=1}^{N} \gamma_k N(\mathbf{x}, \mathbf{m}_k, \mathbf{\Sigma}_k) \tag{8}$$

Here γ_k are the mixture weights. $N(\,)$ is a n-variate Gaussian distribution, and $\mathbf{m}_k$ and $\mathbf{\Sigma}_k$ are the corresponding mean vector and covariance matrix of dimension $n \times 1$ and $n \times n$ respectively.

Assume we have two utterances utt_a and utt_b from the two different speakers. Using the Maximum A-Posteriori (MAP) adaptation with utt_a and utt_b from UBM we generate two speaker models $p^a(\mathbf{x})$ and $p^b(\mathbf{x})$ as follow,

$$p^a(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k N(\mathbf{x}, \mathbf{m}_k^a, \mathbf{\Sigma}_k^a) \quad \text{and} \quad p^b(\mathbf{x}) = \sum_{k=1}^{N} \beta_k N(\mathbf{x}, \mathbf{m}_k^b, \mathbf{\Sigma}_k^b) \tag{9}$$

The distance between these two distributions is defined using the KL divergence [7],

$$KL(p^a \,||\, p^b) = \int p^a(\mathbf{x}) \log(\frac{p^a(\mathbf{x})}{p^b(\mathbf{x})}) d\mathbf{x} \tag{10}$$

We denote $KL(\boldsymbol{\alpha} \,||\, \boldsymbol{\beta})$ as the KL divergence between two probability mass functions $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_N\}$ and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_N\}$

$$KL(\boldsymbol{\alpha} \,||\, \boldsymbol{\beta}) = \sum_{k=1}^{N} \alpha_k \log(\frac{\alpha_k}{\beta_k}) \tag{11}$$

As log is a convex function, the following important result can be obtained [8],

$$KL(p^a \,||\, p^b) \le KL(\boldsymbol{\alpha} \,||\, \boldsymbol{\beta}) + \sum_{k=1}^{N} \alpha_k KL(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) \tag{12}$$

Here we can make some simplification. It is evident from the above equations that, at the time of MAP adaptation, if we do not adapt the mixture weights (i.e. $\alpha_i = \beta_i$ for all i) then $KL(\boldsymbol{\alpha} \,||\, \boldsymbol{\beta}) = \sum_{k=1}^{N} \alpha_k \log(\frac{\alpha_k}{\beta_k}) = 0$. Therefore we have following simplified result,

$$KL(p^a \,||\, p^b) \le \sum_{k=1}^{N} \alpha_k KL(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) \tag{13}$$

Here we have the following closed-form solution of the right hand side,

$$KL(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_k^b|}{|\boldsymbol{\Sigma}_k^a|} + \frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}_k^a (\boldsymbol{\Sigma}_k^b)^{-1}) - \frac{n}{2} + \frac{1}{2} (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (\boldsymbol{\Sigma}_k^b)^{-1} (\mathbf{m}_k^a - \mathbf{m}_k^b)$$

We use the symmetrized version of the KL divergence as follows,

$$KLS(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) = KL(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) + KL(N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b) \,||\, N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a))$$

$$\Rightarrow KLS(p^a \,||\, p^b) \le \sum_{k=1}^{N} \alpha_k KLS(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) \tag{14}$$

$$KLS(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_k^b|}{|\boldsymbol{\Sigma}_k^a|} + \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_k^a|}{|\boldsymbol{\Sigma}_k^b|} + \frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}_k^a (\boldsymbol{\Sigma}_k^b)^{-1}) + \frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}_k^b (\boldsymbol{\Sigma}_k^a)^{-1}) - n$$

$$+ \frac{1}{2} (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (\boldsymbol{\Sigma}_k^b)^{-1} (\mathbf{m}_k^a - \mathbf{m}_k^b) + \frac{1}{2} (\mathbf{m}_k^b - \mathbf{m}_k^a)^t (\boldsymbol{\Sigma}_k^a)^{-1} (\mathbf{m}_k^b - \mathbf{m}_k^a)$$

$$= \frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}_k^a (\boldsymbol{\Sigma}_k^b)^{-1}) + \frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}_k^b (\boldsymbol{\Sigma}_k^a)^{-1}) - n + \frac{1}{2} (\mathbf{m}_k^a - \mathbf{m}_k^b)^t ((\boldsymbol{\Sigma}_k^a)^{-1} + (\boldsymbol{\Sigma}_k^b)^{-1}) (\mathbf{m}_k^a - \mathbf{m}_k^b) \tag{15}$$

Here we simplify the problem farther. We assume, at the time of MAP adaptation only the mean vectors are adapted. Therefore the covariance matrices are same for both the adapted speaker models. Specifically we have, $\boldsymbol{\Sigma}_k^a = \boldsymbol{\Sigma}_k^b = \boldsymbol{\Sigma}_k$

Therefore, the above equation becomes as follow,

$$KLS(N(\mathbf{x}, \mathbf{m}_k^a, \boldsymbol{\Sigma}_k^a) \,||\, N(\mathbf{x}, \mathbf{m}_k^b, \boldsymbol{\Sigma}_k^b)) = \frac{n}{2} + \frac{n}{2} - n + \frac{1}{2} (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (2\boldsymbol{\Sigma}_k^{-1}) (\mathbf{m}_k^a - \mathbf{m}_k^b)$$

$$= (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (\boldsymbol{\Sigma}_k^{-1}) (\mathbf{m}_k^a - \mathbf{m}_k^b) = \mathrm{d}^2(\mathbf{m}_k^a, \mathbf{m}_k^b) \tag{16}$$

With the help of above simplification, the symmetrized version of KL divergence between two speaker models is bounded by the following inequality,

$$KLS(p^a \parallel p^b) \le \sum_{k=1}^{N} \alpha_k (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (\mathbf{\Sigma}_k^{-1})(\mathbf{m}_k^a - \mathbf{m}_k^b) = \mathrm{d}^2(\mathbf{m}^a, \mathbf{m}^b)$$

Here $\mathbf{m}^a = [(\mathbf{m}_1^a)^t\ (\mathbf{m}_2^a)^t \cdots (\mathbf{m}_k^a)^t]^t$ and $\mathbf{m}^b = [(\mathbf{m}_1^b)^t\ (\mathbf{m}_2^b)^t \cdots (\mathbf{m}_k^b)^t]^t$

Therefore from the above inequality it is evident that if the distance between $\mathbf{m}^a$ and $\mathbf{m}^b$ is small, the corresponding $KLS(p^a \parallel p^b)$ is also small.

Using the polarization identity, from the above distance function the corresponding inner product can be calculated as follows,

$$\begin{aligned}
\mathrm{d}^2(\mathbf{m}_k^a, \mathbf{m}_k^b) &= \left\|\mathbf{m}_k^a - \mathbf{m}_k^b\right\|^2 = (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (\mathbf{\Sigma}_k^{-1})(\mathbf{m}_k^a - \mathbf{m}_k^b) \\
&= (\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^a - (\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b - (\mathbf{m}_k^b)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^a + (\mathbf{m}_k^b)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b \qquad (17) \\
\Rightarrow \left\|\mathbf{m}_k^a + \mathbf{m}_k^b\right\|^2 &= (\mathbf{m}_k^a + \mathbf{m}_k^b)^t (\mathbf{\Sigma}_k^{-1})(\mathbf{m}_k^a + \mathbf{m}_k^b) \\
&= (\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^a + (\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b + (\mathbf{m}_k^b)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^a + (\mathbf{m}_k^b)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b \\
\left\langle \mathbf{m}_k^a, \mathbf{m}_k^b \right\rangle &= \frac{1}{4}\left(\left\|\mathbf{m}_k^a + \mathbf{m}_k^b\right\|^2 - \left\|\mathbf{m}_k^a - \mathbf{m}_k^b\right\|^2\right) = \frac{1}{4}(2(\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b + 2(\mathbf{m}_k^b)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^a) = (\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b
\end{aligned}$$

In a similar way the overall distance between the two models can be represented with the help of inner product which is the kernel function of SVM.

$$\begin{aligned}
\mathrm{d}^2(\mathbf{m}^a, \mathbf{m}^b) &= \sum_{k=1}^{N} \alpha_k (\mathbf{m}_k^a - \mathbf{m}_k^b)^t (\mathbf{\Sigma}_k^{-1})(\mathbf{m}_k^a - \mathbf{m}_k^b) = \sum_{k=1}^{N} \alpha_k (\mathbf{m}_k^a)^t \mathbf{\Sigma}_k^{-1} \mathbf{m}_k^b \\
&= \sum_{k=1}^{N} (\sqrt{\alpha_k}\, \mathbf{\Sigma}_k^{-\frac{1}{2}} \mathbf{m}_k^a)^t (\sqrt{\alpha_k}\, \mathbf{\Sigma}_k^{-\frac{1}{2}} \mathbf{m}_k^b) = K(utt_a, utt_b) \qquad (18)
\end{aligned}$$

5 Utterance Partitioning in SVM Classifier with GMM Supervector

The detail discussion of utterance partitioning and training of SVM classifier with GMM supervectors is given below:

We consider an N class problem. In the usual training process (i.e. training without utterance partitioning) from the training speech of the hypothesized speaker, the sequence $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ of feature vectors is extracted. All these feature vectors are used to adapt the UBM. After adaptation, the GMM supervector is created for the speaker. This is same for all the speakers. Hence from the training utterance of each speaker, we extract the sequence of feature vectors and used all these feature vectors to adapt the UBM and create one GMM supervector for that speaker. It is obvious that, in the usual training process (i.e. training without utterance partitioning) from each speaker we will get only one supervector.

As it is a N class problem, we have to create N number of SVM models. At the time of training of speaker one, we use the supervector from speaker one as positive

class data and all the supervectors from remaining $N-1$ classes as the negative class data. Similarly at the time of training of speaker two, we use the supervector from speaker two as positive class data and all the supervectors from remaining $N-1$ classes (i.e. class one, class three,…, up to class N) as the negative class data. Therefore, in the usual training process, the positive class will have only one training sample and the negative class will have $N-1$ training samples. However, in case of SVM classifier training with utterance partitioning we proceed as follows:

From the training speech utterance of the speaker, we calculate the sequence $\mathbf{X}=\{\mathbf{x}_1,\cdots,\mathbf{x}_T\}$ of feature vectors and divide the sequence into P number of subsequences. Using each subsequence we adapt the UBM and create one GMM supervector for that speaker. Therefore from P number of subsequences, we will get P number of GMM supervectors for each speaker. At the time of training of speaker one, we use the P number supervectors from speaker one as positive class data and all the $NP-P$ supervectors from remaining $N-1$ classes as the negative class data.

As an example, we consider 120 second speech utterance from each speaker and the value of $N=100$ and $P=3$. Therefore for each speaker we will have three supervectors. First supervector corresponds to the feature vectors of training speech from initial 40 seconds only. Second supervector corresponds to the feature vectors of training speech from 40 seconds to 80 seconds only. Third supervector corresponds to the training speech from last 40 seconds only. Therefore from 100 speaker class we will have total 300 supervectors. At the time of training of speaker one, we use the 3 supervectors from speaker one as positive class data and all the 297 supervectors from remaining 99 classes as the negative class data.

6 Experimental Evaluations

We have used the POLYCOST database [5] for experimental evaluation of the performances of GMM-UBM classifier and SVM classifier with GMM supervectors. It is evident from recent literatures that MFCC is the most popular feature set for speaker recognition. Therefore we have also used the MFCC feature set [9, 10] for comparative evaluation of the two classifiers. We have used 20 filters. After DCT the first coefficient (i.e. DC value) is discarded since it contains only the energy of the spectrum and the resulting 19 dimensional vector is used.

6.1 Database Description

The POLYCOST database [5] was recorded as a common initiative within the COST 250 action during January-March 1996. It contains around 10 sessions recorded by 134 subjects from 14 countries. Each session consists of 14 items. The database was collected through the European telephone network. The recording has been performed with ISDN cards on two XTL SUN platforms with an 8 kHz sampling rate. Four speakers (M042, M045, M058 and F035) are not included in our experiments as they provide sessions which are lower than 6. All speakers (130 after deletion of four speakers) in the database were registered as clients.

For training the speaker model, we have concatenated the speech form first five sessions and used to train classifier. All the data for each speaker from session six to last available session for an individual speaker were used for testing.

6.2 Preparation of GMM-UBM Classifier

In the present work, we have created a single gender-independent UBM. From the 130 speakers of the POLYCOST database which we have prepared, we collected the speech of 15 male speakers and 15 female speakers. We pooled the feature vectors from all the 30 speakers and trained the 512 mixture UBM using EM algorithm. Diagonal covariance matrices were used for training the GMM.

Remaining 100 speakers were used for training and testing. To build the final speaker model, for each speaker, we have used 2 minutes of training data to adapt the UBM. Only the mean vectors of the UBM have been adapted. For adaptation we choose the relevance factor, $r = 16$.The test data has been tested with all speakers. Hence for each test utterance, there will be one true score and 99 false scores. For testing we have used three types of test speech length. They were 20 seconds, 10 seconds and 5 seconds long. A single, speaker independent, threshold is swept over the two sets of scores and the probability of miss and probability of false alarm are computed for each threshold. Finally we have calculated the equal error rate (EER).

6.3 Preparation of SVM Classifier with GMM Supervectors

We train the SVM classifier with GMM supervector linear kernel. Initially we have not used utterance partitioning. Therefore complete 120 seconds of training speech were used to adapt the UBM. Only mean vectors were adapted and as discussed earlier the GMM supervectors were produced for each speaker. Concisely we can say, after adaptation of UBM model, take the mean vector of any Gaussian and multiply with square root inverse of corresponding diagonal covariance matrix of that Gaussian. The resultant vector is then scaled up by corresponding weight of that Gaussian. This process is repeated for all 512 Gaussians. Lastly we concatenated all these vectors and form a single vector called the GMM supervector. In our case of 19 dimensional feature vector and 512 mixtures UBM the size of each GMM supervector is 9728. We have prepared the SVM classifiers using 1 versus rest way. At the time of testing, we first prepared the GMM supervector using test utterance with MAP adaptation of UBM. After that, apply this test supervector to the SVM model to calculate the score.

6.4 Comparison of GMM-UBM Classifier with SVM Classifier Using GMM Supervector

Table 1 shows the equal error rate performance between the GMM-UBM classifier and SVM classifier with GMM supervector.

Table 1. Comparison of equal error rate (EER (%)) between GMM-UBM base line and SVM with GMM supervector

	Test Speech Lengths		
	20 seconds	10 seconds	5 seconds
GMM-UBM baseline	7.19	7.38	7.97
SVM with GMM super-vector	5.98	9.18	17.90

It is clearly evident from Table1 that, for large test utterance (20 seconds) the performance of SVM classifier is better compared to the GMM-UBM base line. However for small test utterance (5 seconds) the performance of SVM classifier is very poor compared to the GMM-UBM baseline.

At the time of preparation of training supervectors, we have used 120 seconds of training speech to adapt the UBM. Therefore the training supervectors are well adapted. However when we use small test utterances to create test supervectors, then there is a mismatch of adaptation between the training supervectors and test supervector. Due to this mismatch, the performance of SVM classifier with GMM supervector is very poor for small test utterance.

6.5 Effect of Utterance Partitioning in SVM Classifier Using GMM supervector

Table 2 shows the effect of utterance partitioning in the equal error rate of SVM classifier with GMM supervectors. It is clearly evident from Table 2 that, utterance partitioning can improve the verification accuracy significantly. For small test utterance (5 seconds) the improvement is quite high. For large test utterance (20 seconds) also there is a small amount of improvement.

Table 2. Effect of utterance partitioning in EER (%) in SVM classifier with GMM supervectors

No of training supervectors /Speaker	Test Speech Lengths		
	20 seconds	10 seconds	5 seconds
2	5.34	6.30	11.31
3	5.27	5.78	8.57
4	5.41	5.75	7.29
6	5.42	5.40	6.21
8	5.34	5.30	5.83
12	5.63	5.57	5.87

It is evident from the Table 2 that, the optimal number of training supervectors per speaker is eight. For a total training utterance of 120 seconds eight supervectors imply that, the length of each sub-utterance is 15 seconds. After that if we do more partitioning, the equal error rate increases. The reason is that, large number of partitions will create very short sub-utterances. Therefore the MAP adaptation of UBM is very small. Hence the training supervectors of all the speakers are very similar and the discrimination capability of the classifier decreases.

7 Conclusions

We compared the GMM-UBM speaker verification system with SVM speaker verification system using GMM supervector as the linear kernel. Results revealed that, for large test utterance the SVM classifier performed better compared to the GMM-UBM system. However for small test utterance the performance of SVM classifier was poor. To overcome the mismatch between MAP adaptation of training GMM supervector and test GMM supervector, an efficient approach called utterance partitioning was discussed. Results revealed that, the utterance partitioning can improve the accuracy of SVM classifier significantly. It was clear from the results, for proper utterance partitioning, the accuracy of SVM classifier for small test utterance was also significantly better compared to the GMM-UBM baseline.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
2. Cristianini, N., Taylor, J.S.: Support Vector Machines. Cambridge University Press, Cambridge (2000)
3. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support Vector Machines using GMM Supervectors for Speaker Verification. IEEE Signal Processing Lett. 13, 308–311 (2006)
4. Mak, M.W., Rao, W.: Utterance partitioning with acoustic vector resampling for GMM-SVM speaker verification. Speech Communication 53, 119–130 (2011)
5. Petrovska, D., et al.: POLYCOST: A Telephonic speech database for speaker recognition. In: RLA2C, Avignon, France, April 20-23, pp. 211–214 (1998)
6. Hsu, C.W., Chang, C.C., Lin, C.J.: A Practical Guide to Support Vector Classification, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
7. Campbell, J.P.: Speaker Recognition: A Tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
8. Do, M.N.: Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models. IEEE Signal Processing Lett. 10(4), 115–118 (2003)
9. Davis, S.B., Mermelsteine, P.: Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Processing ASSP-28(4), 357–365 (1980)
10. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication 54(4), 543–565 (2012)

L1 Bengali Phonological Interference on L2 English - Analysis of Bengali AESOP Corpus

Shambhu Nath Saha and Shyamal Kr. Das Mandal

Center for Educational Technology, Indian Institute of Technology
Kharagpur, India
shambhuju@gmail.com, sdasmandal@cet.iitkgp.ernet.in

Abstract. Every language has its own phonemic system, which holds unique as well as common features. A language shares some phonemes with other languages, but no two languages have the same phonemic inventory. Contrastive analysis is the field of study in which different phonemic systems are laid side by side to find out similarities and dissimilarities between the phonemes of the languages concerned. The purpose of this study is to derive which phonemes are used by the L1 Bengali speakers to recognize American English phonemes which are new and similar to their L1 Bengali phonology. The results of this study showed typical phonological problems of American English pronunciation by L1 Bengali speakers which will help to develop Computer Assisted Spoken Language Learning (CASLL) tool for faster acquisition of American English language speaking of L1 Bengali speakers.

Keywords: Contrastive analysis, Phonological interference, Language acquisition, Language transfer, Pronunciation, Phonemes, Phonology.

1 Introduction

English is an international language for communication and its importance grows continuously day by day throughout the world. In large number of countries, English is studied and spoken as second language, so understanding the range of variation present in the English spoken in the world today is fundamental issue for the development of English language education as well as spoken language science and technology. Asia is also an important market for English language education, and it is important to learn about Asian language speakers' English and identify their features. Therefore, the Asian English Speech cOrpus Project (AESOP) was launched in order to construct a common shared large scale English speech corpus of Asian language speakers, and participating institutions in various Asian countries are collecting non-native (L2) English speech corpus of speakers of their official languages using the common recording platform [1]. Huge research in speech science is required to implement the linguistic finding into the development of language pedagogy for second language acquisition. Since India is a multilingual country we have so many variety of Indian English based on the local language and dialect. Combining native and

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 790–798, 2013.
© Springer International Publishing Switzerland 2013

nonnative speakers, in India more people who speak or understand English than any other country in the world. Thus research in Indian English dialects from a multidisciplinary perspective is urgently needed to address issues in communication, learning and technology.

Due to the importance grows day by day, it is necessary to acquire English language properly for second language learner. Language acquisition is one of the most important and fascinating aspects of human development. Various language variables are involved in the language process like phonetics, phonology, vocabulary, morphology, syntax, paralinguistic, pragmatics and discourse. Experience, exposure and age of onset of language learning (critical age) are important factors of second language learning [2]. Pronunciation is one of the important parts of second language learning which involves in production of correct sound in target language. Language structure, stress, rhythm influence the pronunciation in target language. A contrastive analysis is made to compare the language structure of native (L1) and non-native (L2) languages [3]. Those structures that are the same in both languages will be positively transferred into the second language (L2) and those that are different will be negatively transferred. The language transfer occurs at both segmental and suprasegmental levels. Contrastive analysis implies that there are three relationships that L1 and L2 phones can have. A phone within the L1 can have one of the three relationships with a phone in L2. According to Flege (1987), an L1 phone can be the same as, similar to or different from an L2 phone. New or different phones are those that are acoustically different from any phone within the L1. Similar phones are those which share some acoustic properties but not all. Phones that are acoustically same are described as same or identical phones. Same phones will be positively transferred from the L1 to L2, results in no problem to the L2 speakers and native like production patterns. Similar and new phones will be more difficult to acquire. L2 speakers use their L1 phonology when they speak L2 language. Effect of L1 phonology on L2 phonology is called phonological interference. L2 speakers' pronunciation is different from that of L1 speakers due to this phonological interference.

Native (L1) Bengali speakers' English is very different from L1 American English speakers due to the phonological interference. From the theory of contrastive analysis, L1 Bengali speakers have problems with new and similar American English consonant and vowel phonemes in both production and perception, which results in incorrect pronunciation. Based on the contrastive analysis, there are some American English phonemes which are new and similar to Bengali phonemic system. The L1 Bengali speakers have problems with these new and similar phonemes in both production and perception. The objective of this study is to find out which phonemes are used by the L1 Bengali speakers to recognize American English phonemes which are new and similar to their L1 phonology.

2 American English and Bengali Phonemes

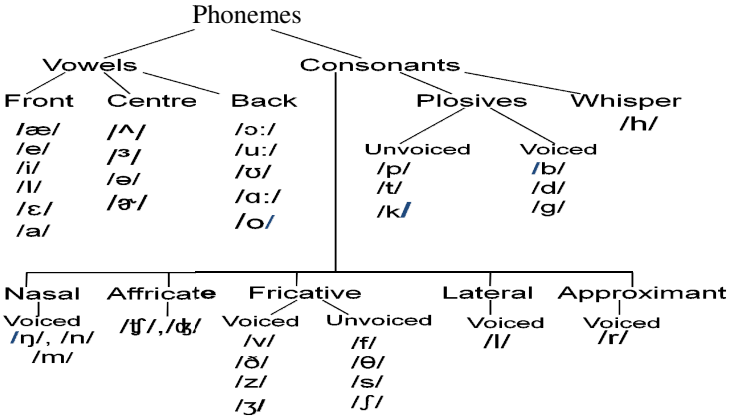


Fig. 1. Phonemes in American English

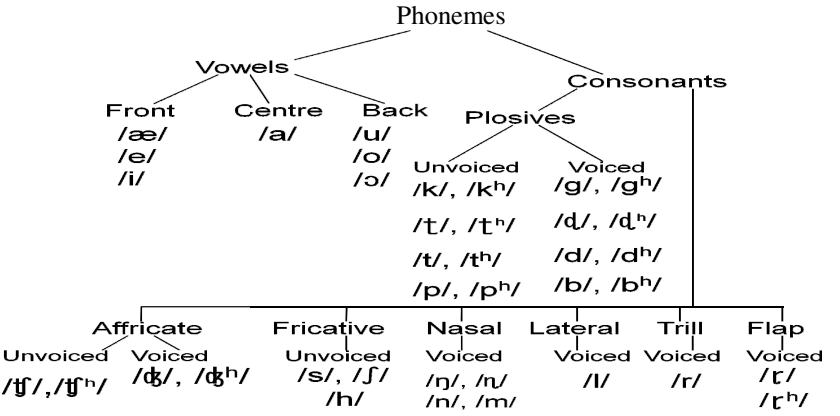


Fig. 2. Phonemes in Bengali

Fig. 1 shows phonemes in American English and Fig. 2 shows phonemes in Bengali. American English has fifteen vowels, where Bengali has a seven vowel system (excluding nasal vowels). On the other hand American English has twenty two consonant phonemes [4], whereas Bengali has thirty one consonant phonemes [5]. Therefore American English has a much larger vowel system than Bengali, but Bengali has more consonant phonemes than American English. Contrastive analysis between consonant phonemes of American English and Bengali represents the same, similar and new consonant phonemes of American English to the L1 Bengali phonology. This type of classification is based on similarities or dissimilarities of place and manner of articulation between American English and Bengali Consonant phonemes. Table 1 represents same, similar and new American English consonants with respect to Bengali consonants.

Table 1. Same, similar, new consonant phonemes to L1 Bengali phonology

American English Phonemes			Bengali Phonemes			Relationship
IPA Symbol	**Place of Articulation**	**Manner of Articulation**	**IPA Symbol**	**Place of Articulation**	**Manner of Articulation**	
k	Velar	Plosive	k	Velar	Plosive	Same
g	Velar	Plosive	g	Velar	Plosive	Same
t	Alveolar	Plosive	t	Alveolar	Plosive	Same
d	Alveolar	Plosive	d	Alveolar	Plosive	Same
p	Bilabial	Plosive	p	Bilabial	Plosive	Same
b	Bilabial	Plosive	b	Bilabial	Plosive	Same
f	Labiodental	Fricative	-	-	-	New
v	Labiodental	Fricative	-	-	-	New
θ	Dental	Fricative	-	-	-	New
ð	Dental	Fricative	-	-	-	New
s	Alveolar	Fricative	s	Alveolar	Fricative	Same
z	Alveolar	Fricative	-	-	-	New
ʃ	Post Alveolar	Fricative	ʃ	Post Alveolar	Fricative	Same
ʒ	Post Alveolar	Fricative	-	-	-	New
h	Glottal	Fricative	h	Glottal	Fricative	Same
ʧ	Post Alveolar	Affricate	ʧ	Post Alveolar	Affricate	Same
ʤ	Post Alveolar	Affricate	ʤ	Post Alveolar	Affricate	Same
m	Bilabial	Nasal	m	Bilabial	Nasal	Same
n	Alveolar	Nasal	n	Dental	Nasal	Similar
ŋ	Velar	Nasal	ŋ	Velar	Nasal	Same
l	Alveolar	Lateral	l	Dental	Lateral	Similar
r	Post Alveolar	Trill	r	Alveolar	Trill	Similar
j	Palatal	Approximant	j	Palatal	Approximant	Same
w	Bilabial	Approximant	w	Bilabial	Approximant	Same

3 Speech Material

The material used for the present study was the Aesop's fable "The North Wind and the Sun", which produces a large range of segmental and suprasegmental characteristics in English. This phonetically balanced passage is recommended by the IPA, which contains 144 syllables, 113 words, 8 independent clauses, 5 dependent clauses, 5 sentences and 3 paragraphs. The material was read by 8 (4 male, 4 female) L2 English (L1 Bengali) speakers whose Native language was Bengali. All speakers were in the age group between 20 to 40 years. The speech was recorded in quiet room directly into a PC, using microphone. For the fluency of reading, the speakers were instructed to read out the text several times before reading and read the material aloud. The speech was digitized at a sampling rate 16 kHz with an accuracy of 16 bits/sample.

4 Methodology

The aim of this analysis is to find out which phonemes are used by the L1 Bengali speakers to substitute American English consonant phonemes which are new (f, v, θ, ð, z, ʒ) and similar (n, l, r) to their L1 phonology and also find out which phonemes are used by the L1 Bengali speakers to recognize American English vowel phonemes. In order to do that, at first the Aesop's fable "The North Wind and the Sun" was phonetically transcribed by using TIMIT dictionary [6]. Then acoustic analysis of speech waveform of every L2 English speaker (L1 Bengali) was performed. In this analysis, spectrogram of every consonant phoneme in L1 Bengali speaker's English speech waveform was verified manually to detect the consonant phoneme correctly and compared it with the corresponding phoneme in L1 American English pronunciation derived from TIMIT dictionary. In case of vowels, measured the F1 and F2 of each phoneme manually from L1 Bengali speaker's English speech waveform to identify the vowel correctly and then compared it with the corresponding phoneme in L1 American English pronunciation derived from TIMIT dictionary. Finally count the number of substitution for each phoneme.

5 Results and Discussions

5.1 Consonants

/ f/

The total number of times consonant /f/ was 40. There were 31 times of /f/ out of 40 times replaced with Bengali consonant phoneme /p^h/ by L1 Bengali speakers in their English speech and remaining 9 times, /f/ was remain same. This result indicates that L1 Bengali speakers substitute /f/ with Bengali phoneme /p^h/.

/v/

The total number of times consonant /v/ appeared was 48. There were 31 examples of /v/ replaced by /b^h/ by L1 Bengali speakers in their English speech and remaining 17 examples, /v/ was remain same. Since number of substitution of /v/ by /b^h/ was higher than the number of not substitution of /v/ (remain same), so the L1 Bengali speakers recognize /v/ as /b^h/.

/θ/

The total number of times consonant /θ/ appeared was 32. All appearances of /θ/ were replaced by /$ʈ^h$/ by L1 Bengali speakers in their English speech. This result indicates that L1 Bengali speakers recognize /θ/ as Bengali phoneme /$ʈ^h$/.

/ð/

The total number of appearance of consonant /ð/ was 200. There were 196 examples of /ð/ replaced by /ɖ/ by L1 Bengali speakers in their English speech and remaining 4

examples, /ð/ was remain same. Since number of substitution of /ð/ by /ɖ/ was higher than the number of not substitution of / ð/ (remain same), so the L1 Bengali speakers recognize /ð/ as Bengali phoneme /ɖ/.

/z/

The total number of appearance of consonant /z/ was 72. There were 8 examples of /z/ replaced by /s/, 13 examples of /z/ replaced by /ʤ/ and remaining 51 times, /z/ was same. This result indicates that L1 Bengali speakers recognize /z/ as /z/ like L1 American English speakers.

/r/

In American English, / ɚ/ is a r-colored shwa and a allophonic variation for the realization of r which follows /ə/. /r/ is not pronounced by L1 American English speakers after / ɚ/, regardless of the position of / ɚ/ in a word. The total number of appearance of consonant / ɚ/ was 88. There were 65 examples where L1 Bengali speakers uttered /r/ after / ɚ/ unlike L1 American English speakers and remaining 23 examples L1 Bengali speakers did not utter /r/ after / ɚ/ like L1 American English speakers. The result indicates that L1 Bengali speakers cannot realize the phonology of English /r/.

/p/, /t/, /k/

In American English, phonemes /p/, /t/, /k/ are aspirated at the beginning of words and at the beginning of stressed-syllables. There were 72 appearances of /t/, out of which 26 appearances were aspirated and 46 appearances were unaspirated. There were 72 appearance of /k/, out of which 4 appearances were aspirated and 68 appearances were unaspirated. There were 35 appearances of /p/, out of which 8 appearances were aspirated and 27 appearances were unaspirated. Since number of unaspirated /p/, /t/, /k/ generated by L1 Bengali speakers was higher than their aspirated counterpart, so it indicates that L1 Bengali speakers cannot realize the phonology related to English phonemes /p/, /t/ and /k/.

/l/, /n/

The total number of appearance of consonant /l/, /n/ was 112 and 208 respectively. For all examples of /l/, /n/, L1 Bengali speakers produced them like L1 American English speakers; only difference is that intensity of / l/, /n/ produced by L1 Bengali speakers were higher than that of L1 American English speakers. This result indicates that L1 Bengali speakers recognize /l/ and /n/ properly like L1 American English speakers.

From above analysis, it is concluded that the L1 Bengali speakers realize new consonant phonemes by substituting their L1 phonemes, which are given in Table2.

Table 2. Substitution of American English phonemes by L1 Bengali speakers

American English	f	v	θ	ð	z	l	n	r
Bengali	p^h	b^h	t^h	ɖ	z	l	n	r

5.2 Vowels

/i:/ , /I/

The total number of appearance of vowels / i:/ and /I/ was 80 and 144 respectively. There were 58 examples out of 80 examples, where / i:/ was replaced by Bengali phoneme /i/ and there were 101 examples out of 144 examples, where / I/ was also replaced by Bengali phoneme /i/. The remaining 22 appearances of / i:/ and 43 appearances of / I/ were not substituted. Since number of substitution of / i:/ and /I/ by /i/ was higher than number of not substitution, so it indicates that L1 Bengali speakers categorize English vowel phonemes / i:/ and /I/ as Bengali vowel phoneme / i/.

/e/ , /ɛ/

The total number of appearances of vowels /e/ and /ɛ/ were 48 and 32 respectively. There were 31 examples out of 48 examples, where /e/ was replaced by Bengali phoneme /e/ and there were 22 examples out of 32 examples, where / ɛ / was also replaced by Bengali phoneme /e/. The remaining 17 appearances of /e/ and 10 appearances of / ɛ / were not substituted. Since number of substitution of /e/ and / ɛ / by Bangla vowel phoneme /e/ was higher than number of not substitution, so it indicates that L1 Bengali speakers categorize English vowel phonemes /e/ and /ɛ/ as Bengali vowel phoneme /e/.

/ æ/

The total number of appearance of vowels /æ/ was 128. There were 88 examples out of 128 appearances, where /æ/ was replaced by Bengali phoneme /æ/; 20 examples of /æ/ were replaced by Bengali phoneme /a/ and remaining 20 appearances of /æ/ were not replaced. Since number of substitution of /æ/ by Bangla vowel phoneme /æ/ was maximum, it indicates that L1 Bengali speakers replace English vowel phonemes /æ/ with Bengali vowel phoneme /æ/.

/a/

The total number of appearance of vowels /a/ was 24. Out of 24 appearances, there were 17 examples where L1 Bengali speakers replaced English /a/ by Bangla /a/ and remaining 7 appearances were not replaced. This result indicates that L1 Bengali speakers replace English vowel phonemes /a/ with Bengali vowel phoneme /a/.

/ʌ/

The total number of appearance of vowels /ʌ/ was 56. Out of 56 appearances, there were 38 examples where L1 Bengali speakers replaced English /ʌ/ by Bangla /a/; there were 4 examples where L1 Bengali speakers replaced English /ʌ/ by Bangla /ɔ/; and remaining 14 appearances were not replaced. Since number of substitution of /ʌ / by Bangla vowel phoneme /a/ was maximum, it indicates that L1 Bengali speakers replace English vowel phonemes /ʌ / with Bengali vowel phoneme /a/.

/ɑ:/

The total number of appearance of vowels /ɑ:/ was 8 and number of substitution of /ɑ:/ by Bangla phoneme /a/ was 6 ; remaining 2 appearances of /ɑ:/ was not

substituted. This result implies that L1 Bengali speakers replace English vowel phonemes /ɑ:/ with Bengali vowel phoneme /a/.

/o/

The total number of appearance of vowels /o/ was 64 and number of substitution of /o/ by Bangla phoneme /o/ was 40 ; remaining 24 appearances of /o/ was not substituted. This result implies that L1 Bengali speakers replace English vowel phonemes /o/ with Bengali vowel phoneme /o/.

/ ɔ:/

The total number of appearance of vowels /ɔ:/ was 112. Out of 112 appearances, there were 72 examples where L1 Bengali speakers replaced English /ɔ:/ by Bangla / ɔ/; there were 8 examples where L1 Bengali speakers replaced English /ɔ:/ by Bangla /o/; there were 3 examples where L1 Bengali speakers replaced English /ɔ:/ by Bangla /a/ and remaining 29 appearances were not replaced. Since number of substitution of /ɔ:/ by Bangla vowel phoneme /ɔ/ was maximum, it indicates that L1 Bengali speakers categorize English vowel phonemes /ɔ:/ as Bengali vowel phoneme /ɔ/.

/u:/, / ʊ/

The total number of appearances of vowels /u:/ and / ʊ/ were 56 and 24 respectively. There were 49 examples out of 56 examples, where /u:/ was replaced by Bengali phoneme /u/ and there were 17 examples out of 24 examples, where /ʊ/ was also replaced by Bengali phoneme /u/. The remaining 7 appearances of /u:/ and 7 appearances of /ʊ/ were not substituted. Since number of substitution of /u:/ and /ʊ/ by Bangla vowel phoneme /u/ was higher than number of not substitution, so it indicates that L1 Bengali speakers categorize English vowel phonemes /u:/ and /ʊ/ as Bengali vowel phoneme /u/.

/ɚ/

The total number of appearance of vowel r-colored shwa /ɚ/ was 88. Out of 88 appearances, there were 50 examples where L1 Bengali speakers replaced /ɚ/ by Bangla /a/; there were 10 examples where L1 Bengali speakers replaced /ɚ/ by Bangla /ɔ /; there were 5 examples where L1 Bengali speakers replaced /ɚ/ by Bangla / æ / and remaining 23 appearances were not replaced. Since number of substitution of /ɚ/ by Bangla vowel phoneme /a/ was maximum, it indicates that L1 Bengali speakers categorize English vowel phonemes /ɚ/ as Bengali vowel phoneme /a/.

/ə/

The total number of appearance of vowel shwa /ə/ was 272. Out of 272 appearances, there were 80 examples where L1 Bengali speakers replaced /ə/ by Bangla /a/; there were 41 examples where L1 Bengali speakers replaced /ə/ by Bangla /ɔ /; there were 26 examples where L1 Bengali speakers replaced /ə/ by Bangla / æ /; there were 21 examples where L1 Bengali speakers replaced /ə/ by Bangla / e / ; 4 examples of /ə/

were replaced by /i/; 4 examples of /ə/ were replaced by /u/ ; 4 examples of /ə/ were replaced by /o/ ;and remaining 95 appearances of /ə/ were not replaced. Since number of substitution of /ə / by Bangla vowel phoneme /a/ was maximum, it indicates that L1 Bengali speakers categorize English vowel phonemes /ə/ as Bengali vowel phoneme /a/. In American English, vowels in unstressed syllables are reduced in quality, intensity, duration and centralized to /ə/ [7]; in Bengali, there is a tendency to try to equalize the duration of each syllable, and the presence or absence of accent does not affect vowel quality or duration [8]. In American English, vowel reduction in unstressed syllable is an important feature. Here it was observed that the vowel /ə/ had the highest number of variants with 177 examples and various vowels were substituted for the shwa /ə/ by L1 Bengali speakers. This indicates that vowel weakening does not occur in L1 Bengali speakers' English and so vowels in unstressed syllables remain strong. That means, vowel reduction is a difficult feature for all L1 Bengali speakers to acquire. From this analysis it is concluded that Bengali speakers categorize American English vowels with respect to Bengali vowel phonemes.

6 Conclusions

The results of this study showed typical phonological problems of English pronunciation by Bengali speakers which have been often discussed in EFL. This analysis showed Bengali phonemes that are used by the L1 Bengali speakers to substitute American English same and new consonant as well as vowel phonemes. The Results of this analysis will assist the development of Computer Assisted Spoken Language Learning (CASLL) tool for faster acquisition of American English language speaking of L1 Bengali speakers.The experiments need more samples and subjects, and also need to be discussed in more detail. In the future study, evaluate the fluency level of the L1 Bengali speakers and derive the relationship between their English pronunciation and fluency level.

References

1. Visceglia, T., Tseng, C., Kondo, M., Meng, H., Sagisaka, Y.: Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project). In: 2009 Oriental COCOSDA, Beijing, China, August 10-12, pp. 52–57 (2009)
2. Saville-Troike, M., Pan, J., Dutkova, L.: Introducing second language acquisition. Cambridge Introductions to Language and Linguistics. Cambridge University Press, Cambridge (1995)
3. Shockey. L, Phonetic and phonological properties of connected speech. Working Papers in Linguistics,17. Columbus: Department of Linguistics. The Ohio StateUniversity (1974)
4. Roach, P.: English Phonetics and Phonology: a practical course, 2nd edn., pp. 120–123. Cambridge University Press (1998)
5. Chatterji, S.K.: The Original and Development of the Bengali Language, pp. 402 and 279 paragraph 3, 3rd impression. Rupa.Co (2002)
6. http://www.ldc.upenn.edu/Catalog/docs/LDC93S1/TIMITDIC.TXT
7. Meng, H., Tseng, C., Kondo, M., Harrison, A., Visceglia, T.: Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper. In: Interspeech, pp. 1715–1718 (2009)
8. Basu, J.B., Mitra, T., Mandal, M., Das, S.K.: Grapheme to phoneme (g2p) conversion for Bangla. In: Oriental COCOSDA International Conference on Speech Database and Assessments (2009)

Evolution of the Modern Phase of Written Bangla: A Statistical Study

Paheli Bhattacharya[1] and Arnab Bhattacharya[2]

[1] Govt. College of Engineering and Textile Technology, Serampore, Hooghly, India
pahelibhattacharya@gmail.com
[2] Dept. of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India
arnabb@iitk.ac.in

Abstract. Active languages such as Bangla (or Bengali) evolve over time due to a variety of issues. In this paper, we analyze the change in the written form of the modern phase of Bangla quantitatively in terms of character-level, syllable-level, morpheme-level and word-level features. We collect three different types of corpora—classical, newspapers and blogs—and test whether the differences in their features are statistically significant. Results suggest that there are significant changes in the length of a word when measured in terms of characters, but there is not much difference in usage of different characters, syllables and morphemes in a word or of different words in a sentence. To the best of our knowledge, this is the first work on Bangla of this kind.

1 Introduction

Bangla (or Bengali) is one of the most widely spoken languages. It belongs to the Indo-European family of languages and is believed to have been derived from Prakrit in around 650 CE. The history of Bangla is divided into three phases: Old Bangla (till 1350 CE), Medieval Bangla (1350-1800 CE) and Modern Bangla (1800 CE-).

Since its inception, Bangla, like any other active language, has undergone a lot of changes due to a variety of social, cultural, economic and political causes. The changes happen mostly in vocabulary and pronunciation, one of the big catalysts for which is the adoption of words of foreign origin either directly or indirectly into the language. For example, in Bangla, there is no word that depicts the concept "football" directly, and consequently, the English word has been adopted verbatim and has become part of the language now. Similarly, the English word "box" has been incorporated in Bangla as বাক্স (bAksa[1]) by suitably modifying its pronunciation.

A particularly remarkable source of variety in Bangla is the two clearly distinct forms of written prose in the modern phase – *Sadhu Bhasha* (chaste language) and *Chalit Bhasha* (colloquial language). The chaste language was used earlier and has been now replaced in almost all communications in Bangla by the colloquial version. The most notable change has happened in the form of verbs and pronouns which has become shorter and can be more easily pronounced. For example, the verb করিয়াছি (kariAChi) has become করেছি (karaChi) and the pronoun তাহাদের (tahadera) has been transformed to তাদের (tader).

[1] We have used the ITRANS transliteration mechanism to specify the words in Bangla font.

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 799–805, 2013.
© Springer International Publishing Switzerland 2013

With the advancement of digital world, the electronic media have imparted a large impact on the modern language which is clearly reflected in newspapers, blogs and social networking forums. It is extremely rare to find longer words such as যৌবনতেজোদীপ্ত (JaubanatejodIpta) now than in the classical literature.

However, while all these notions of change are commonly believed to be true, to the best of our knowledge, there is no work that tests whether these perceptions about the differences are *statistically significant*. In this paper, we precisely aim to fill this gap. Our main contribution, thus, is to study the changes in the modern phase of written Bangla in a statistically robust manner.

We collect three different corpora – one consisting of classical literature, and the other two that of newspapers and blogs (the details are in Section 3). We then extract different features at the word and sentence levels and test whether the changes across the corpora are significant when viewed from a statistical standpoint.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3, Section 4 and Section 5 describe the corpora, the features and the statistical testing method respectively. Section 6 discusses the results before Section 7 concludes.

2 Related Work

Evolutionary linguistics or the study of evolution of languages has long fascinated human beings. In addition to numerous studies that have been developed, there are whole conferences—EvoLang, the Evolution of Language International Conferences (`evolang.org`)—that are devoted for this. Among the computational studies for language evolution and change, an overwhelming majority of the work is focused on European languages [8,3,6].

Indian languages, due to the relative paucity of digital resources, had not been studied deeply. The recent ease of using Unicode and the surge of excellent work in the field of natural language processing (NLP) have, however, changed the situation dramatically. Sikder analyzed the change in type of words in Bangla and showed how the frequency of foreign words are increasing [7]. Choudhury et al. studied the change of Bangla verb inflections for the single verb কর (kara), which means “to do” in English [2]. They gave a functional explanation for the rise in several dialects of Bangla because of phonological differences while uttering the verb inflections.

To the best of our knowledge, computational studies describing the changes in the Bangla language, however, have not been undertaken so far. The evolution of a language is most visible when changes occur in the discourse. Since we are still far off from that for Bangla, in this paper, we study some micro-level aspects of the written language in terms of characters, syllables, morphemes and words.

3 Corpora

For our work, we collected three different corpora:

1. *Classical Corpus*: It includes the literary works of 4 eminent authors.

Table 1. Number of words in the corpora

Corpus	Classical	Newspaper	Blog
Number of words	1,58,807	7,71,989	5,18,485

2. *Newspaper Corpus*: It includes the news articles from 7 leading newspapers of both India and Bangladesh.
3. *Blog Corpus*: It includes blog articles (but not the comments) from 11 blogs.

The details are in the full version of this paper [1]. The total number of words in each of the corpus are listed in Table 1.

4 Features

4.1 Character-Level Features

In Bangla, there are two types of characters—vowels and consonants. The consonants cannot be pronounced on their own and must always end with the sound of a vowel. Vowels, on the other hand, can be pronounced on their own and are written either as independent letters or as diacritical marks on the consonant they attach to. For example, ক্ (k) is a consonant. When it is joined with আ (A), it is written as কা (kA). Thus, the diacritical mark for the vowel আ (A) is া (A^2). The vowel অ (a) has an invisible diacritical mark. Its only effect is to remove the ্ (the consonant-ending marker) from the consonant it attaches to. Thus, ক্+অ (k + a) is written as ক (ka).

We distinguish the diacritical mark of a vowel from the vowel itself as the latter can stand on its own. For example, the correct parsing of খুশীতে (khushIte) is খ্+ু+শ্+ী+ত্+ে (kh+u+sh+I+t+e) and that of আলোক (Aloka) is আ+ল্+ো+ক্+· (A+l+o+k+a) where · is used to represent the invisible diacritical mark of the vowel অ (a). The four consonants, ৎ, ং, ঃ, ঁ (t.h, .n, H, .N respectively), are treated differently in that they do not have the consonant-ending marker ্. Thus, বাংলা (bA.nlA) is parsed as ব্+া+ং+ল্+া (b+A+.n+l+A).

Conjunct characters where two (or three) consonants are joined together are parsed differently. There is no vowel at the end of the first (respectively, the first two) and only the last consonant has a vowel ending written as a diacritical mark. Hence, the correct parsing of সন্ত্রস্ত (santrasta) is স্+·+ন্+ত্+র্+·+স্+ত্+· (s+a+n+t+r+a+s+t+a).

Character Frequencies. We count the frequencies of all the characters—consonants, vowels and diacritical marks—using the parsing system discussed above for the three corpora. The number of distinct characters is 61 that includes the 39 consonants (the consonant ব (b) is counted only once), the 11 vowels (the vowel ঌ (no ITRANS code) is not used any more) and the corresponding 11 diacritical marks.

We also count the frequencies of bi-gram and tri-gram characters. For example, the bi-grams in the word বাংলা (bA.nlA) are বা (bA), াং (A.n), ংল্ (.nl) and লা (lA). The tri-grams are extracted similarly.

[2] The ITRANS coding for the diacritical marks remain the same.

We arrange the uni-gram characters (and bi-grams and tri-grams) in descending order of their frequencies. When comparing corpus C_1 with C_2, we consider the top-50 entries from the sorted list of C_1 and find their frequencies in C_2. Thus, the comparison of C_1 with C_2 differs from that of C_2 with C_1 as, in the later case, the frequencies of the top-50 entries of C_2 are considered. The frequencies from the two corpora form the two non-parametric distributions between which the changes are statistically tested. Instead of using the raw counts as frequencies, we compute the relative ratios by dividing by the total number of characters in the corpus; this makes two corpora of differing sizes comparable.

Character-Based Word Length. For each corpus, we produce a count of words that have a particular length in terms of characters. Thus, if there are 300 words of length 4, the frequency corresponding to 4 in the non-parametric distribution is 300. The distribution consists of all the word lengths and their frequencies. The comparison of corpus C_1 with C_2 is symmetric for this feature.

4.2 Morpheme-Level Features

A *morpheme* is the smallest meaning-bearing unit in a language. A morpheme may not be able to stand on its own, although a word necessarily does. Every word is composed of a root word (sometimes called a *lexeme*) and possibly one or more morphemes.

To extract morphemes, we used the unsupervised program Undivide++[3], based on [5]. Unfortunately, the program have many parameters, and even after repeated tuning and discussion with the authors, we could not replicate the accuracies as reported in [5] on our corpora for the 4110-word test-set provided by them. Although the program is only about 50% accurate on average, we still use it to extract all the morphemes from the words in the corpora. (The full version of this paper [1] reports the performance on the metrics as proposed in [4]).

Morpheme Frequencies. For every morpheme, we get a count of words that have it. Similar to the character frequencies, we then extract the top-50 (normalized) frequencies from each corpus.

Morpheme-Based Word Length. Using the program Undivide++, every word is segmented into a list of prefix(es), root word and suffix(es). The "length" of the word is then counted as the number of such segments. For example, if প্রদেশটিকে (pradeshaTike) is segmented into the prefix প্র (pra), the root দেশ (desha) and the two suffixes টি (Ti) and কে (ke), its length is counted as 4.

4.3 Syllable-Level Features

Syllables are the smallest subdivisions uttered while pronouncing a word. Since syllables are phonetic units, they cannot be extracted completely correctly without speech analysis. To bypass the problem, we employ a very simple and intuitive heuristic.

[3] Available from `http://www.hlt.utdallas.edu/~sajib/Morphology-Software-Distribution.html`

Table 2. Features used

Feature type	Level			
	Character	**Syllable**	**Morpheme**	**Word**
Uni-gram frequency	yes	yes	yes	yes
Bi-gram frequency	yes	yes	no	yes
Tri-gram frequency	yes	no	no	no
Length of word or sentence	yes	yes	yes	yes

We assume that any combination of characters till the next vowel is a syllable. Thus, each vowel, each consonant with its vowel ending (encoded as a diacritical mark), and each conjunct character is a separate syllable.

The consonants, ৎ, ং, ঃ (t.h, .n, H respectively), are treated as single syllables since they do not have the consonant-ending marker ্. However, ঁ (.N) is considered part of the preceding syllable. Thus, অকস্মাৎ (akasmAt.h) has three syllables অ (a), ক (ka), স্মা (smA) and ৎ (t.h) while বাঁধা (bA.NdhA) has two syllables বাঁ (ba.N) and ধা (dhA).

Syllable Frequencies. For every uni-gram syllable (and also bi-grams of syllables), we get a count of words that have it. We again consider only the top-50 (normalized) frequencies from each corpus.

Syllable-Based Word Length. Similar to characters, the word length is also counted in terms of syllables.

4.4 Word-Level Features

The words are parsed from the sentences using orthographic word boundaries (i.e., the white-space characters including ?, !, . and the Bangla character ।).

Word Frequencies. The words are for sentences what the characters are for words. Thus, this feature is computed in exactly the same way as characters.

Word-Based Sentence Length. Similar to word length, the sentence length is counted in terms of number of words.

5 Statistical Testing

All the features that are used in this paper are summarized in Table 2.

To test whether the distributions of the various features for the different corpora are statistically different from each other, we employ the non-parametric two-sample *Kolmogorov-Smirnov (K-S) test*. For each pair of corpora, we perform three tests. Suppose the corpora are C_1 and C_2. The *null hypothesis* H_0 for all the three tests state that the samples observed empirically for C_1 and C_2 come from the *same* distribution.

There can be three ways by which the *alternate hypothesis* can vary. For the non-equal ($\neq$) test, the alternate hypothesis $H_A^{\neg}$ states that the empirical values $x_i^{(1)}$ and

$x_i^{(2)}$ for the distributions from C_1 and C_2 are different, i.e., for every i, $x_i^{(1)} \neq x_i^{(2)}$. For the greater than ($>$) test, the alternate hypothesis $H_A^{>}$ states that the empirical values for C_1 are greater than the corresponding values for C_2, i.e., for every i, $x_i^{(1)} > x_i^{(2)}$. The less than ($<$) test is similar where the alternate hypothesis $H_A^{<}$ tests whether for every i, $x_i^{(1)} < x_i^{(2)}$.

The K-S test returns a *p-value* that signifies the confidence with which the null hypothesis can be rejected. The lower the p-value, the more statistically significant the result is. Thus, for the $H_A^{\neg}$ case, it means the two distributions are more different. If the result of a $\neq$ test is statistically significant at a particular level of significance, then the result of either the $>$ test or the $<$ test (but not both) must be significant as well at the same level of significance.

6 Results

The differences between the word lengths in terms of number of characters between the three corpora are found to be statistically significant[4] for the alternate hypothesis $H_A^{\neg}$. (The tables in the full version of this paper [1] list all the p-values.) More interestingly, the alternate hypothesis $H_A^{<}$ is found to be very significant for classical versus blog, classical versus newspaper and blog versus newspaper comparisons. This shows that the frequency of words having a shorter length is less in classical than in blogs which, in turn, is less than newspapers. Thus, this shows that longer words were more common in the classical literature than in newspapers which are more than that in blogs.

Although the classical corpus exhibits longer words in terms of syllables (due to the $H_A^{<}$ test), the non-equality test ($H_A^{\neg}$) is not significant. This, thus, indicates that the use of conjunct characters were more in classical literature which led to longer words in terms of characters but not in terms of syllables. The blogs and newspapers differ in terms of number of syllables though.

The differences in number of morphemes is again not significant. Thus, contrary to popular perception, words with many suffixes and prefixes are not more abundant in the classical literature as compared to the current scenario. Similarly, the number of words per sentence for classical is not statistically different either.

Frequencies of uni-gram characters, bi-gram characters and uni-gram syllables are not significantly different across the corpora. Frequencies of tri-gram characters, bi-gram syllables, uni-gram words and bi-gram words of classical are significantly different from both blogs and newspapers for the alternate hypotheses $H_A^{\neg}$ and $H_A^{<}$. The newspaper and blog corpora show little statistical difference in frequencies indicating that the current styles of formal and informal writing are quite alike.

7 Conclusions

In this paper, we provided a model of statistically testing the differences of writing styles across various phases of a language. To the best of our knowledge, this is the first work of its kind in Bangla. This work has aimed at building a basic foundation on

[4] Unless otherwise mentioned, we consider the level of significance to be 5%.

which more analysis in terms of higher-level features can be carried out in the future. Also, bigger corpora will allow robust and more detailed analyses of the results.

References

1. Bhattacharya, P., Bhattacharya, A.: Evolution of the modern phase of written Bangla: A statistical study. arXiv [cs.CL] (2013)
2. Choudhury, M., Jalan, V., Sarkar, S., Basu, A.: Evolution, optimization, and language change: The case of Bengali verb inflections. In: ACL SIG Computational Morphology and Phonology, pp. 65–74 (2007)
3. Christiansen, M.: Language evolution. Oxford University Press (2003)
4. Dasgupta, S., Ng, V.: Unsupervised morphological parsing of Bengali. Language Resources and Evaluation 40(3-4), 311–330 (2006)
5. Dasgupta, S., Ng, V.: High-performance, language-independent morphological segmentation. In: HLT-NAACL, pp. 155–163 (2007)
6. Niyogi, P.: The Computational nature of language learning and evolution. MIT Press (2006)
7. Sikder, S.: Contemporary bengali language. Amor Ekushey (February 21, 2013), http://archive.thedailystar.net/suppliments/2013/Amor%20Ekushey%20Speci%al%20Supplement/pg2.htm
8. Steels, L.: The synthetic modeling of language origins. Evolution of Communication 1(1), 1–34 (1997)

Contextualizing Time in Linguistic Discourse: Cues to Individuate and to Order Events

Samir Karmakar

School of Languages and Linguistics, Jadavpur University, India
samirkrmkr@yahoo.co.in

Abstract. Temporal description of a discourse is a consequence of the interactions holding among different contextualities. These interactions are governed by different contextualization cues of both explicit and implicit types. The explicit contextualization cues are tense, grammatical aspect, connectives etc., whereas the implicit contextualization cues are expectancies and dependencies emerging from the concept internal structures of open class expressions in a description. In this paper, an attempt would be made to understand how some of these cues play a crucial role in construing the sense of temporality encoded in a discourse, with an example drawn from *Bangla*.

Keywords: landmark, affordances, foreground, background, temporality.

1 Introduction

Last few decades have witnessed a growing concern among researchers about the issues of subjectivity relevant to temporality and temporal ordering; since real world ordering is changed or modified by the language user to reflect his or her subjective preferences [6, 14]. This concern is developed in two ways. (i) For one group of researchers, the time flow of discourse remains the central issue. The object of their investigations is the linguistic expressions larger than a single sentence and the issue of sequentiality is dealt with in terms of successive bounded events [2]. (ii) For others, the issues of attention and attitude in constructing a discourse are of focal interest [16, 19]. As per this view, a discourse and its interpretation are shaped by our understanding of the world. This has an obvious consequence on the use of language in discourse.

This paper investigates a *Bangla* discourse as a paradigm to explore the issues of time flow in discourse following the proposal of the first group of researchers, and also try to understand the role of attention and attitude while construing a discourse. What will remain implicit in this analysis is the fact that, in a communicative context the interaction between different explicit textualities with subjective intentions constructs the temporal view of the discourse [5, 7, 8, 15]. Subjective intentions, as the 'metacommunicative functions', are crucial in invoking the context. Gumperz [9] identifes these metacommunicative functions as 'contextualization cues', a subset of discourse markers in his framework.

A framework for the study of the narrated discourse is outlined in Section 2. In Section 3 different contextualization cues relevant to the temporal interpretation of a

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 806–813, 2013.
© Springer International Publishing Switzerland 2013

discourse are discussed with special reference to *Bangla* data followed by a conclusion in Section 4.

2 The Framework

A temporal description represented in a discourse will be examined from the perspectives of procession and function to see how temporal order is construed out of the contextualities.

2.1 Procession: Landmark and Affordances

In a description, some of the situations have explicit textual correspondences and some do not. The set of all such explicit and implicit situations belongs to the class we designate as *landmark* class. A landmark represents a state of affairs in terms of the participating individuals, relations and spatio-temporal locations. The interdependencies of the participants are represented schematically. There is already a body of literature in the field of semantics dealing with such schematic representation at the level of sentence [12, 17]. In contrast to the sentential level, discourse level representations deal with the interconnected sentences. Each sentence comes with a new piece of information, hence figuring a distinct aspect of the context. The interrelations between the sentences cause continuous modification of the existing state of knowledge, establishing links between different parts of the context. We need to consider how sub- and supra-sentential interactions are governed by the general structure of presupposed knowledge commonly shared by the interlocutors. Further, implicit landmarks are named as *affordances*. Affordances are useful in capturing various types of inferential roles which hardly have any explicit realizations.

2.2 Function: Foreground and Background

The function of the landmarks in a discourse could be classified as either *foreground* or *background*. Those landmarks which directly contribute to the basic construction of the description are to be considered as having foregrounding significance. On the other hand, those states of affairs which elaborate the basic construction or the skeleton of the discourse are of backgrounding significance. Foregrounded landmarks play a crucial role in conceptualizing the dynamics of the description, hence contributing to the sequential ordering of the corresponding situations. In contrast, backgrounded landmarks of a description amplify or elaborate the basic construction of the discourse [2, 10], hence are crucial for the simultaneous ordering of the situations.

3 Time in Discourse

3.1 Temporal Organization of Discourse

Interpreting discourse level temporal relations often calls for processing the socio-cultural presuppositions which the users of a language invoke during their interaction through the characteristic use of linguistic expressions. These presuppositions

constitute the notion of 'shared knowledge', and are functional in understanding the communicative intention of a narration [11]. Consider the following narrative:

(1) S_1: gato SonibAr AmAr jibon-er Saroniyo din ch-il-o
last Saturday my life-of memorable day be-past-3_{past}
Last Saturday was a memorable day in my life.
S2: Ami jathA riti nodi-te SnAn kor-te giy-ech-il-Am
I as usual river-in bath do-for go-perf-past-1_{past}
I had gone to river for bathing, as usual.
S3: SnAn-er ghAT-e takhon bhiR ch-il-o nA
of-bathing ghat-in then crowded be-past-3_{past} not
The bathing ghat was not crowed then.
S4: kebol du-Ti chele Sei Somoy nodi-te SnAn
only two-cl boy that time river-in bath
kor-ch-il-o
do-impf-past-3_{past}
Only two boys were bathing in the river that time.
S5: tAder modhye ek-jon gobhir jal-e poR-e
them among one-cl deep water-in fall-part
SAhAjj-er jonno chitkAr kor-e
help-of for cry do-pres-3_{press}
Having fallen into the deep water one of them cries for help.
S6: Ami tA Sun-te pA-i
I that hear-part get-pres-1_{pres}
I hear that.
S7: tAr-par Ami tAk-e uddhAr kor-i
that-after I him rescue do-pres-1_{pres}
After that I rescue him.

Generation and interpretation of temporality encoded in (1) seeks to explore the roles of speaker and hearer in terms of their pragmatic significances. Scholars investigating the time flow of the discourse have been less concerned about the role of speaker and hearer as two important constituents of a discourse. As has been pointed out by Binnick [2], their interest revolves around the notion of reference time shift. To this group of scholars, the temporal organization of a discourse is relative to the order of the situations. The ordering of the situations is governed by the semantics of nouns, verbs, adverbs, connectives, tense, aspect etc. These ordered situations are the basis of the internal time of the macro-situation. On the other hand, for the other group of researchers, the speaker and the hearer are two important constituents of the discourse [1].

Information encoded in (1) is enumerated below. The enumeration constitutes a class of landmarks. Some of it is functional in foregrounding, some in backgrounding the situations. In addition, it can also invoke the necessary inferences in order to construct a coherent temporal description of a discourse to fill the "narrative gap".

(2) Inferences:
S_1: Information about the *orientation* and *evaluation*
(L_{1a}) the temporal location (orientation)
(L_{1b}) speaker's attitude to the incident (evaluation)

S_2: Information about *orientation*, *evaluation* & *complicating action*
(L_{2a}) the spatial location (orientation)
(L_{2b}) speaker's habit of bathing regularly in the river (evaluation)
(L_{2c}) going event of the speaker to the river (complicating action)
(L_{2d}) reaching event of the speaker (complicating action)
S_3: Information about *orientation*
(L_3) the location where the incident took place in terms of the individuals other than the speaker
S_4: Information about *complicating action* and *orientation*
(L_4) the individuals, other than the speaker, who were at that time in the location of the incident
S_5: Information about complicating *action* and *orientation*
(L_{5a}) the falling event of an individual into the water (complicating action)
(L_{5b}) the depth of the water (orientation)
(L_{5c}) the cry for help by an individual (complicating action)
S_6: Information about *complicating action*
(L_6) the hearing event about the cry for help
S_7: Information about *complicating action*, and also *result/resolution*
(L_7) the rescuing event of an individual

Following Labov [13], we have classified above list of landmarks in terms of their respective roles, like *abstract*, *orientation*, *evaluation*, *complicating action*, *result/resolution* and *coda*.[1] As per this analytical framework, a single sentence can have more than one information type. Structural roles of the landmarks in (2) have certain intentional imports, which are crucial in determining the use of temporally relevant linguistic symbols.

3.2 Foregrounded Landmarks in the Temporal Description

A careful investigation into the landmarks enumerated above would reveal that L_{2d}, L_6 and L_7 are indispensible for the description of the incident. Out of these three landmarks, L_{2d} is not explicitly mentioned in the text. It is inferred from S_2, where the verb form, representing "volitional movement form one place to another by an agent", being inflected with past tense and perfect aspect resolves into a sense of reaching. This inference also imposes a mutual entailment on L_{2c} and L_{2d}, that is $L_{2c} \models L_{2d}$ and vice versa.

But how does one consider these three landmarks as the core constituents of the same temporal description? - The coherence of a whole description depends on (a) the

[1] Abstract is given to provide an unbiased objective interpretation of any incident. To keep the narrative tension intact, here in (1) the technique of abstracting is not adopted by the narrator. Orientation introduces information relevant to characters, temporal and physical settings and situations. The evaluation expresses the attitudinal stance of the narrator to the narrated situation. The complicating action delineates the sequence of the situations leading to the climax. The result/resolution informs what finally happens, hence releases the tension which is built by the complicating action. Coda announces the end of the narration.

identity relation holding among individuals across the landmarks, (b) the causal relation holding between the successive situations, and (c) the temporal shift in construing the perspective.

(a) Identity:

In a first person narrative like (1), the speaker acts as a frame of reference and also as a connecting thread to define the macro-situation external and macro-situation internal times respectively. The macro-situation external time concerns the A-series time, since it locates the macro-situation with respect to the speaker's position. On the other hand, the macro-situation internal time is about the B-series time, since micro-situations are ordered with respect to one another. As a consequence, in a first person narrative the role of speaker "I" is to provide a frame of reference, while construing the macro-situation external time. But to construe the macro-situation internal time, the speaker "I" provides a thread to connect the foregrounded landmarks.

In *Bangla*, the identification of the speaker can be done on the basis of the person marker of the inflected verb forms and pronouns: the individual in S_2 who went to the river, the individual in S_6 who heard the cry for help and the individual in S_7 who rescued the victim are identical. The identification of the speaker across the sentences integrates the corresponding landmarks into a thread. In addition, since the individuals are in space time, the identity relation also helps to fix the speech time. In isolation, S_2, S_6 and S_7 have different speech times, but under identification they reflect a single speech time.

(b) Causal Structure:

This section will discuss those causal schemes of our temporal cognition which are crucial in foregrounding the flow of time because of affording the sense of sequentiality. Those schemes of causation are enumerated below:

(3) If A, B, C are three situations, then
  a. (A *causes* B) → (A *precedes* B)
     Example from S_5,
     (falling causes crying) → (falling precedes crying)
  b. (A causes B) and (B precedes C) → (A precedes C)
     Example from $S_{5,7}$,
     (falling *causes* crying) & (crying *precedes* rescuing) → (falling *precedes* rescuing)
  c. (A *causes* B) and (B *overlaps* with C) → (A *precedes* C)
     Example from $S_{5,6}$,
     (falling *causes* crying) & (crying *overlaps* with hearing) → (falling *precedes* hearing)

The concept of causation is defined as irreflexive, non-commutative, and transitive. Irreflexive property states that A cannot be caused by itself. As per noncommutative property, if A causes B, B cannot cause A. If this is not the case, then the conceptual scheme stated in (3a) will collapse. The transitive property states if A causes B and B causes C, then A causes C. This conceptualization of causation is crucial in imposing a sense of temporal direction, and therefore becomes one important factor in temporal flow expected by the foregrounded landmarks.

(c) Temporal Shift:
The problem arises with the use of tenses in case of S_2, S_6 and S_7. In S_2 the sentence is marked with the past tense, whereas S_6 and S_7 are marked with present tense. As a consequence, the reference time of L_{2d} precedes speech time, whereas in L_6 and L_7 the corresponding reference times and speech time overlap with each other. How is it possible to identify the individuals when they are temporally dislocated? – Answer to this question would not obviously come from the semantics itself, since the use of present tense in case of narrating a past incident is a matter of subjective attitude towards the incident [18].

The foregrounded landmarks constitute the order of the macro-situation internal time. Relative positioning of the corresponding situations projects the temporality internal to the macro-situation, which is of the B-series type, since the foregrounded landmarks are ordered with respect to each other.

(4) $\underbrace{L_{2d} < L_6 < L_7}_{\text{Internal Ordering}}$

In contrast, the external ordering sets up a relation between the utterance time (represented as "U") and the time of the macro-situation, and therefore, is of A-series type. While fixing the external ordering, the macro-situation is viewed as a "whole". Time external to the macro-situation is concerned about its location in past, present or future with respect to the speech time (U). The complete temporal structure of the discourse in (1) would be as follows:

(5) $\underbrace{\underbrace{L_{2d} < L_6 < L_7}_{\text{Internal Ordering}} < U}_{\text{External Ordering}}$

3.3 Backgrounded Landmarks in the Temporal Description

In contrast to L_{2d}, L_6 and L_7, other landmarks play their background role by providing following information to the foregrounded core represented in (5): (a) a referential system and (b) a connecting system. Both these systems are crucial in processing the information relevant to abstract, orientation, evaluation, resolution, and coda. I will discuss the functioning of these two systems with a special reference to the adverbial and pronominal forms used in example (1) to investigate the role of backgrounded information in construing the temporal interpretation of a discourse.

(a) Adverbials: Studies in the Referential System
Adverbials in a discourse need to be classified on the basis of the *frame-of-reference* they often require while indicating a point in time. Three different frames of reference are proposed here: (i) *ego-centric* (ii) *event-centric* and (iii) *allo-centric*.

In the light of the above mentioned frames of reference, the adverbials of (1) will be discussed here. The adverbial use of *gato SonibAr* "last Saturday" in S_1 locates the entire happening, with respect to the narrator's position on the time line. Decoding the phrase "last Saturday" needs to employ the semantics of the adverbial such as "last"

and the knowledge about the clock-calendar time. The adverb “last” when uttered by the narrator refers to sometime which is situated in the past, and not within the span of narrator's present time. On the other hand, “Saturday” behaves like an ordinal since it denotes a position within the period of the week. Therefore, decoding the information represented in `last Saturday' invokes both ego-centric and allo-centric referencing.

Though the day of the incident has been mentioned, the narration gives no explicit information about the exact time of the happening. The time of the incident is, however, derived from S_2. The use of the adverbial form *jathA riti* “as usual” here in this context entails a sense of being “used to”. When one does something as usual, it entails that (s)he is habituated to perform some action periodically over time [18]. Like “used to”, “as usual” has a sense of modality because of conveying a sense of regular necessity; hence, it presumes a sense of daily routine. The time of the incident should be inferred with respect to that daily routine of the narrator. Moreover the use of *jathA riti* “as usual” in S_2, in contrast to the use of the adjectival form *Saroniyo* “memorable” in S_1, intensifies the distinction between the events of the “mundane” life and the events of significance. The significance of an event in memory is often perceived in the backdrop of a mundane life experience. This technique maximizes the impact of the description in narration.

(b) Pronominals: Studies in the Connecting System
Pronouns are important in maintaining the integrity of a discourse as a coherent temporal whole. Discourse pronominals behave like discourse tense, in the sense that interpretation in both the cases is relative to the notion of reference.

In (1), five pronouns are employed by the narrator: (i) *takhon* “then” in S_3, (ii) *Sei Somoy* “that time” in S_4, (iii) *tAder* “among them” in S_5, (iv) *tA* “that” and (v) *tAr-par* “after that” in S_7. These five pronouns are crucial in interrelating the events and individuals in the narrative. The pronouns mentioned in (i), (ii), (iv) and (v) connect two different event times, whereas (iii) specifies an individual.

Both *takhon* “then” and *Sei Somoy* “that time” refer to the time of bathing. *tAder* “them” refers to the two boys bathing in the river. *tA* “that” refers to the cry of the drowning boy for help. *tAr-par* is interpreted with respect to the hearing of the drowning boy's cry. These pronominal forms integrate the information into a coherent whole.

4 Conclusion

While investigating the way time is contextualized in (narrated) discourse, this paper investigates different contextualization cues, under the assumption that cohesion or the connectivity in a discourse can be deduced from the dependencies found at the lexical level. The interconnections holding between these lexemes in terms of their inter- and intra-sentential expectancies and dependencies finally yield a coherent temporal description irrespective of their (a)temporal natures as it follows from Brown and Yule [3] also.

References

1. Bach, K., Harnish, R.N.: Linguistic communication and speech acts. MIT Press, MA (1979)
2. Binnick, R.I.: Time and the verb: A Guide to Tense and Aspect. Oxford University Press, New York (1991)
3. Brown, G., Yule, G.: Discourse Analysis. Cambridge University Press, Cambridge (1983)
4. Cresswell, M.: Static Semantics for Dynamic Discourse. Linguistics and Philosophy 25, 545–571 (2002)
5. Dowty, D.: The effects of Aspectual Class on the Temporal Structure of Discourse: Semantics or Pragmatics? Linguistics and Philosophy 9, 37–61 (1986)
6. Fleischman, S.: Evaluation in Narrative: The Present Tense in Medieval "Performed Stories". Yale French Studies 70, 199–251 (1986)
7. Grice, P.: Studies in the Way of Words. Harvard University Press, MA (1989)
8. Grimes, J.E.: The thread of discourse. Mouton Publishers, The Hague (1975)
9. Gumperz, J.: Discourse Strategies. Cambridge University Press, Cambridge (1982)
10. Hopper, P.J.: Some observations on the typology of focus and aspect in narrative language. Studies in Language 3(1), 37–64 (1979)
11. Ifantidou, E.: Evidentials and Relevance. John Benjamins Publishing Company, Amsterdam (2001)
12. Jackendoff, R.: Semantic Structure. MIT Press, MA (1990)
13. Labov, W.: The transformation of experience in narrative syntax. In: Languages in the Inner City. University of Pennsylvania Press, Philadelphia (1972)
14. Longacre, R.E.: The Grammar of Discourse. Plenum Press, NY (1983)
15. Nerbonne, J.: Reference Time and Time in Narration. Linguistics and Philosophy 9, 83–95 (1986)
16. Östman, J.O., Virtanen, T.: Discourse Analysis. In: Handbook of Pragmatics Manual. John Benjamins Publishing Company, Amsterdam (1995)
17. Pustejovsky, J.: The Generative Lexicon. MIT Press, MA (1995)
18. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of The English Language. Longman, London (1985)
19. Stalnaker, R.C.: Context and Content. Oxford University Press, NY (1999)

Utterance Discourse and Meaning: A Pragmatic Journey with the Bangla Discourse Particle /na/

Rimi Ghosh Dastidar[1] and Sibansu Mukhopadhyay[2]

[1] School of Languages and Linguistics, Jadavpur University, Kolkata
rimigd@gmail.com
[2] Society for Natural Language Technology Research, Kolkata
sibansu@gmail.com

Abstract. For some cases meaning only depends on the manner of utterances. Discourse particles as the discourse-marking expressions are such cases which are extremely valuable in a discourse and exceptionally ambiguous in terms of language processing. This paper provides a particular case study with the elaborated examples of typical context sensitivity of the utterances of Bangla discourse particle *na*. We also attempt to analyze syntactic distribution. Such as *topicalization* and *tag question* formation with *na* in Bangla. Each type of *na* is exemplified with a spectrum graph to indicate the variation of the *utterances*. This attempt follows a descriptive framework of pragmatic analysis of discourse.

Keywords: Utterance, topicalization, tag question. Illocutionary act.

1 Introduction

In a given context, collaboration between more than one set of expressions preceded by the necessary presuppositions in the participating agents' minds make a conversation total. In an ordinary conversation presuppositions help a speaker to produce utterances and a listener to understand the expression. The participants in this case should be well aware of the standard expressions of the language through which they are collaborating the dialogue in the given context. This awareness is as usual considered as pragmatic competence.

This paper evaluates some cases when the participating agents in a conversation are eventually considered as pragmatically competent in Bnagla and the agents know how to use *na* in the several contexts rather than the postulated version of the Bangla negatives. Therefore the basic problem comes in this paper is to disambiguate several *na*s in Bangla used in the several context and these are not to be considered as negative. These non negative *na*s can be considered as discourse particles as discussed in the generative literature since last a few decades. Let us unfold the problem with some examples cited in (1) and (2)

(1) (a) tumi *na* aj ratTa ekhanei theke jao. (*"You rather stay here tonight."*)
You *na* today night-cl here-i-emp stay-cv-imp-2nd-fut

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 814–822, 2013.
© Springer International Publishing Switzerland 2013

(b) Shyamaler *na* mathaY kono buddhi nei. (*Shyamal has no sense at all)*
Shyamal *na* head-in any sense not
(c) Mohar Ekhono boiTa pOReni *na*? (*"Has not Mohor yet read the book?")*
Mohor still book-cl read-3rd –past-not *na*?
(d) tumi ajkei ciThiTa likhe phElo *na*. (*"You write the letter today"*)
You today-i-emp letter-cl write-cv-imp-2nd-fut *na*
(e) Minai aj kOthaTa boluk *na*
Mina today word-cl speak-subj-3rd *na*
"Let Mina speak it out today."

Every example cited in (1) contains *na* without negation contrasts with the *na*s cited in the example (2). *Na*s occur in (1) are the discourse particles.

This paper points out pragmatic variation of those discourse particles which occur also as the negative elements as usual in the languages. We provide a comparative parallel data set from Bangla in this paper. We also refer the pragmatic contexts of the discourse particle *na* in Bangla. Approaching negation *na* in Bangla always occurs with a verb either following or in the preceding place. Conventional format of sentential negation in Modern Bangla is V+NEG, although in conditionals and in the subjunctives the order can be seen as NEG+V. Both of V+NEG and NEG+V is exemplified in (2).

(2) (a) ami bhat khabo na (*"I shall not eat rice"*)
I rice eat-1-fut not
(b) ami jodi bhat na khai SEm rag korbe
I if rice not eat Shyam angry will-be
(c) ami cai o bhat na khak *("I want him not to have a tea")*
I want-1-pre she rice not eat-subj

In older Bangla NEG+V was the usual format, which is changed into V+NEG order 'due to a real pressure of neighboring non Indo-Aryan Languages.' (Singh 1976, cited in Banerjee 2001) However in modern Bangla sentential negation is expressed by a single negative head *na* or *ni* (cf. Dutch negation, cited in Haegeman and Zanuttini, 1991) and also by the phenomenon called negative concord. Negative in Bangla cannot occur in the PRO infinitive clause if not condition applied. For example consider (3) and (4),

(3) SEmol [PRO ciThiTa likhte] caY na
Shyamal [PRO letter-cl write-inf] want-3-pre not
(4) * SEmol [PRO ciThiTa likhte na] caY
*Shyamal [PRO letter-cl write-inf not] want-3-pre

Chatterji (1970) argued that the insertion of *na* between noun and verb was a very common process in middle Bangla. For instance, *sahan na jaY* "not be tolerated". In modern Bangla *na* is inverted after verb, such as *sahan jaY na* "it cannot be borne". But *na sahana jay* is not found in any older form of Bangla. It is supposed that *na* played the role of quantifier of verb at that time. Therefore, *na* cannot be separated from a verb by the noun. (Chatterji, 1970: 925) But in Modern Bangla we find construction like (5),

(5) cakrir jonno kothaY *na* ghurechilo KOmol
Job for where *na* roam-3-past Kamal

In Chatterji 1970, we also notice another example in different context *eso na* "do come in". According to Chatterji this *na* is derived from Sanskrit *naama* as in "*aviSata naama*". (1970: 520) He has left a clue only saying that this *na* is not a negative element rather this is "*nirbhandha na*".

In Indo Aryan languages, /nO/-avyaya occurred before the verb till old and middle Bangla. As a consequence, it was used as prefix with a number of verbs and it gradually generated negative verbs being added/ combined with those verbs in middle Bangla. /nOho/ is such a negative verb. In middle Bangla /nOho/ is used in all tenses. Such as in present /nOhe/, past /nOhilo/, and in future /nOhibo/, etc. In this period we also find /naMro/, /nare/, /narie/, /narilo/, /naribo/. We can notice negative verb in Srikrisnakirtan but they are not single unit but pre negative combined verb, such as 'nade' (<na + de = dEY na(.) Sen, 2007: 259)

2 *na* as a Discourse Particle

A simple distinction between utterance and speech should be clarified first to very lightly demonstrate the title of this paper. Utterance is basically mentioned as the taxonomy of the discourse particle is strongly associated with the spoken form of language. It is considered as a smallest unit of speech with a version that has to be interpreted by the pragmatic competence of the native speakers and hearers whereas speech is designated with the established grammatical characteristics. A normal discourse, for example as perpetuated daily conversation happens, is an amalgamation of utterances with effective meanings presupposed in the speaker-hearer's cognitive domain. *na* as a discourse particle in Bangla can prove all these things to be taken the idea granted.

For many cases, the word *na* in Bangla in spite of being a negative, shows several "positive" appearances in the environments where it occurs. The case of where a straight grammatical category has turned into a lexically meaningless but pragmatically valuable unit is supposed to be a case of *grammaticalization*. The central claim of *grammaticalization* process is that the 'change in grammar is gradual rather than discrete.' and this process is a unidirectional process. (Harris 1997) Therefore, the hypothetical reconstruction for a *grammaticalized* component is very difficult although we have given some hints on the position of *na* in the historical stages of Bangla language in section 1. Though we do not find evidences through which we can clearly denote the deviation of *na* from negation to the function as discourse particle in Bangla, we may generalize that it is also a case of *grammaticalization*.

The major tasks of this paper are then to describe pragmatically the variation of Bangla discourse particle *na* subject to the various contexts with exemplification and elaborated spectrum graphs to indicate the variation of the utterances.

3 Illocutionary Act and Modal Character

Performative utterances as introduced in Austin (1975) are considered as illocutionary act of speaking. These utterances may be or may not be performed as the well-defined linguistic categories shape a conversation towards a discourse. Illocutionary act does not describe a reality passively but it forces to change a real condition to another. So that the intended meaning works as the direct stimulus to the hearer's mind. This force also works to shape a sentence into the intended modal projections of the speaker. We shall now describe some of these shapes as showing the variations to be considered as the Discourse-Syntax-Interface phenomena.

Discourse particle *na* in Bangla is a modal exponent which works in the various pragmatic situations and generates a special attitude of the speaker in terms of the discourse. It shows diversity in its function. Illocutionary force of a phrasal articulation can only be launched by a discourse particle. *na* also attests such force to the phrasal content. *na* as the non-negative discourse particle occurs in the various modal conditions. A few examples cited in below:

- **Declarative:** In declaratives *na* occurs mostly after an NP.

(6) ami *na* aj jabo na.("I am not going today.")
I na today go-1st-fut-not

- **Interrogative:** *na* can form interrogative sentences forming tag-questions. This particular use of *na* can be focused as a desire of a positive answer.

(7) ar ekTuo chocolate nei, *na*? ("there is no more chocolate, right?")
(8) ajker chobiTa beS bhalo chilo, *na*?("Today's movie was great, isn't it?")

- **Imperative:** *na* occurs in the imperative sentence both after noun and verb.

(9) Amake ektu jOl dao na ("Give me a glass of water.")
(10) Tumi na ekbar oder baRi jao.("Please go to their house once.")

It is really very interesting to observe that the presence of discourse particle changes a sentence with immense possibility. The sentences exemplified in the above being contained that of the discourse particle *na* are as if look like the bifurcated parts of the total discourses. As if every sentence is plucked and detached from the contexts and giving hints more to say about the conditionality of a supposed conversation. So that the utterances with the illocutionary force with which speakers correlate hearers. We have seen that DSI phenomenon like topicalization or tag-question formation is very often observed when a conversation is conditioned to be a discourse with the natural utterances. On the other hand, these particular DSI matters are considered as the very well-know phenomena in the late 80's generative traditions, namely root-sentence or root-clause phenomena or root-phenomena in short.

4 Topicalization and Tag-Questions

4.1 Topicalization

One of the very crucial observations here is that the uses of *na* as discourse particle in Bangla are seen mainly in the root clause. For example, *na* exhibits *topicalization* and

tag-question formation in Bangla. An important function in such case of *na*/ is that it sometimes works as a topic marker in Bangla like another Bangla discourse particle *to*.

(12) ei boiTa *to* sOkoler pORa ucit
this book-cl *to* everybody-gen read should "This book everybody should read"
Like *to* used in (12), *na* is also able to mark or separate a topic in a construction. Consider following examples (13), especially (13)(a) which is provided just like a replica of (12):
(13) (a) ei boiTa na sOkoler pORa ucit
this book-cl na everybody-gen read should "This book everybody should read"

Consider more examples to observe the cases of marking:

(b) TomaY na Jadob cinte pare ni. "Jadav does not recognize you."
(c) Tomake-i na jadob cinte pare ni. baki sObaike pereche.
(d) * ami mone kori na je tomaY na jadav cinte pare ni
(e) * tomaY na jadav cinte pare ni bole ami mone kori na.

In (13) topic marker /na/ contains a prosodic peak that falls upon "tomaY" or "tomake" comparable to "ei boiTa" and predicate accent marks the focus of the construction. It conceives a gap in the clause and that moves the noun phrase to the initial position of the sentence and attaches NP to the root S. This is a root sentence phenomenon. (Emonds, 1976: 30-31) Embedding is not possible for this type of construction. So that we cannot accept either (13)(d) or (13)(e).

4.2 Tag-Questions

Tag-question has a tag shown in (14). Important thing is that the subject of the tag clause is the antecedent of the subject in main clause. In English tag has the opposite value from the main clause. But In Bangla, both negative and positive statement in the main clause followed by /na/.

(14) He loves you, doesn't he?
Similarly consider (15),
(15) (a) SEmol Simulke bhalobaSe, na? "Shyamal likes Shimul, doesn't he?"
(b) SEmol Simulke bhalobaSe na, na? "Shyamal doesn't like Shimul, does he?"

5 Context-Sensitivity

We have randomly chosen five sentences each of which has *na* as a discourse particle. Simultaneously for the experiment we made five replicas of the chosen sentences where intentionally we have deleted theses *na*s. We ask a Bangla native speaker to articulate the sentences having and without having *na*s for the purpose of recording as we have observed that utterance of discourse particle *na* makes a sentence highly

context-sensitive. Now we move to analyze the pragmatic behavior of Bangla *na* in the different contexts with few sentences simultaneously having a *na* and without a *na*, for showing *na* occurring in the same position can differ in meaning and how *na* itself adds an extra essence to the sentence where it occurs. Each of (A), (B), (C), (D) and (E) is showcasing a pair of sentences which shows contrast for +na and –na in the various contexts. We provide a model of spectrum analysis for each sentence to make our discussion observationally more adequate.

Now in a pilot level study to show sound pattern of the variation of /na/ in different sentential models experimentally we here present spectrum graph for two pairs of sentences. We hope that in this way we can have an overview on the behavior and effect of /na/ when they occur in sentences. The following sentences study how intonation matters in speech. In short, each picture shows "sound pressure level/Hz" in the vertical axis and "Frequency" in the horizontal.

(A)

(i) ami *na* aj ar kOlej sTriT jabo na
I *na* today again college street go-1-fut not

(ii) ami aj ar kOlej sTriT jabo na
I today again college street go-1-fut not

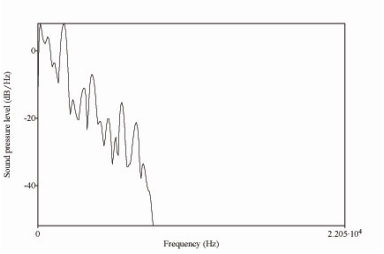


Fig. 1. Declarative with *na*

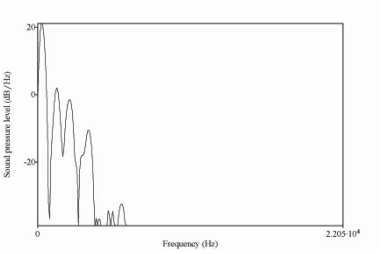


Fig. 2. Declarative without *na*

The speaker says somebody that he is not willing to go to College Street. If we withdraw first 'na' from the sentence, the meaning of the sentence does not alter. But insertion of *na* between the pronoun and verb adds a kind of politeness to the sentence and changes the tone of the speaker from declaring just a statement to expressing a desire. This may mean or the one projected pragmatics of this sentence (i) may be that the speaker intends to overlay her tiredness onto the hearer's mind for not going to the college street. (i) and (ii) both are to state that the speaker wants to break the regularity of going to college street, whereas (i) is modifying the approach of speaker. The sentences are normally negative and declarative in terms of modality. The terminal *na* indicates the negation. But the surprising thing to notice is two *na*s, identical in form but they are doing completely different function. Consider (B) and (C) for more variations, may be intended to increase politeness of the utterances, but specifically not to elaborate tiredness as the possibility of (A)(i).

(B)

(i) tui *na* khub bhalo meye!
You *na* very good girl!

(ii) tui khub bhalo meye!
You very good girl!

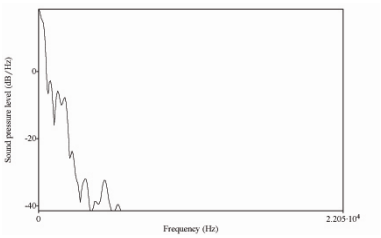


Fig. 3. Declarative Exclamative with *na*

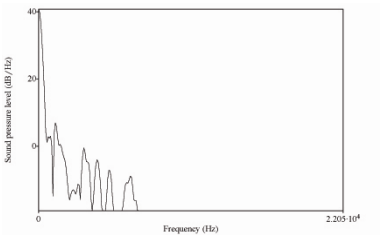


Fig. 4. Declarative Exclamative without *na*

With (B) (i) The speaker says to a girl that she is really good. The speaker is assured of her goodness and with a caring and loving tone asserts this. Here the *na* increases the softness of the speaker's voice. The sentence (B) (ii) without *na* is just a mere statement. The speaker admires a girl with a statement which contains no extra ordinary tonal quality. The second sentence is like 'yes, you are very good girl'.

(C)

(i) tui *na* khub bhalo meye?
You *na* very good girl?

(ii) tui ki khub bhalo meye?
Are you a very good girl?

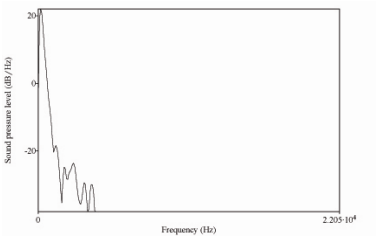


Fig. 5. Declarative Interrogative with *na*

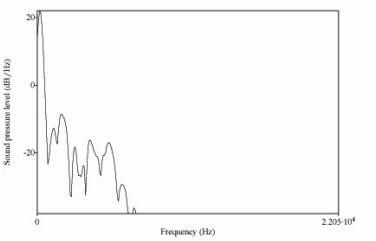


Fig. 6. Declarative Interrogative without *na*

The mode of both (C)(i) and (C)(ii) is question. In the first one the speaker tries to convince the girl something by saying that she is a good girl. *na* adds a kind of strength to the sentence. /na/ here works both as particle as well as a question marker where second sentence too is a question with a normal question marker.

(D)

(i) tui ca khabi na, na?
You tea drink not, *na*?

(ii) tui ca khabi na?
You tea drink not?

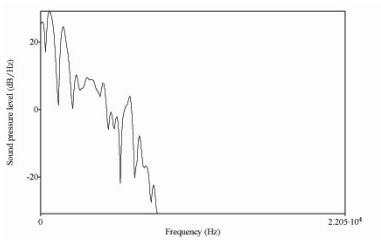


Fig. 7. Tag-Question

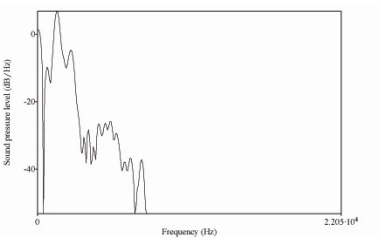


Fig. 8. Simple Questioning

(D(i) and D (ii) both are question regarding tea taking. From the both sentences we can realize that in both cases the speaker is already aware of second person's unwillingness for tea. But in first one the speaker wants to be confirmed by asking somebody whether the second person is not really eager to have a tea. 'tui ca khabi na' is a negative sentence and terminal 'na' is a particle and adds an extra assurance both from the speaker and listener. The second one of the pair is a mere question where the speaker asks someone whether he won't have tea?

(E)

(i) PhonTa dhOr na!
Phone-cl receive *na!*

(ii) PhonTa dhOr
Phone-cl receive

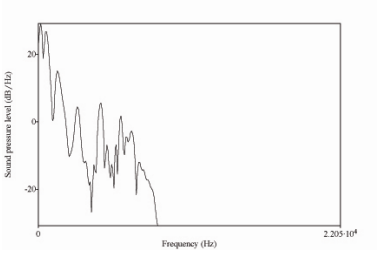


Fig. 9. Imperative with *na*

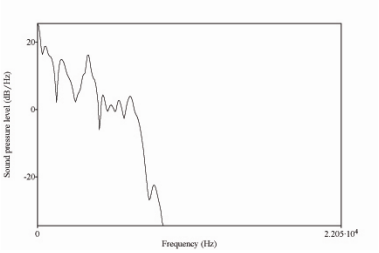


Fig. 10. Simple Imperative

We can suppose here that somewhere phone is ringing continuously and the speaker is busy with some work and being disturbed with the monotonous sound of the phone she tells somebody in the room to pick up the phone. If she expresses herself with (E)(i) there lies an emphasis in her tone and if she utters (E) (ii) it will be just a flat imperative, non emphatic one without containing any additional quality.

The diagrams display diverse graph for each sentences. In one hand as we find different intonation for identical sentence including and excluding /na/ and notice how dynamically /na/ works according to context. If we notice the graphs carefully an obvious rise and fall of the curve can be located in the diagram in form of peaks. Whenever we insert /na/ within the sentences the curve line ranges high at the particular point where *na* is posited. So if we compare the sentences in one pair, the sentence with /na/, to be specific, the position of discourse particlae *na* shows a greater amplitude at the same frequency than the other.

6 Conclusion

From the natural instinct of language one can draw the similarity between the previous and present form of a word. It is to remember here it is not necessary that the past and present form share the common meaning. Some words lose their original meaning totally due to evolution of language or some coin altered meaning over the change of time. The total loss of meaning is called bleached out. Many South Asian languages show this feature of meaning loss. Particles are like that. We have noticed two kind of particle in language. One is purely particle and they have no other func-

tion to play in language, such as /go/, /re/. They occur after verb and set a direct relation with the pronoun./to/ is a particle in Bangla, belonging to this category, appear in different pragmatic context and is confined to its one and only role. Apart from having the peculiarities of a particle, they are just without any semantic engagement. Particle can access its meaning only through its context. The other type of particle is different from the previous. They sometimes behave as particle though they have a particular lexical meaning in language. They can occur independently, for instance, /na/./na/ generally is a negative word in Bangla. In this case, it is easy to find a diachronic meaning of /na/. But when /na/ is used as a particle then a kind of void or emptiness is observed within its meaning. Therefore particle /na/ and negative /na/, though identical in physical structure differ completely in their meaning, usages, behavior and distribution. Now the question may arise that the native speakers of Bangla are aware of the fact but what the nonnative speakers recognize the particle /na/ and negative /na/?

References

1. Harris, A.C.: Remarks on Grammaticalization. In: Butt, M., Kings, T.H. (eds.) Proceedings of the LFG 1997 Conference (1997)
2. Bannerji, S.: Grammar of Case Marking and Adpostions: A Parametric Study. PhD Dissertation (2001)
3. Bayer, J., Obenauer, H.-G.: Discourse Particles, Clause Structure, and Question Types, Cambridge, October 30-31 (2008)
4. Bayer, J., Dasgupta, P.: Emphatic Topicalization and the structure of the left periphery: Evidence from German and Bangla (Unpublished) (2011)
5. Brinton, L.J.: Pragmatic Markers in English: Grammaticalization and discourse functions. Mouton de Gruyter, Berlin (1996)
6. Chatterji, S.: The Origin and Development of the Bengali Language (3 vol.). George Allwin, London (1970)
7. Chomsky, N.: Lectures on Government and Binding: The Pisa Lectures. Mouton de Gruyter (1981/1993)
8. Dasgupta, P.: Chinno kOthaY SajaYe tOroni, Kolkata. Gangchil Publication (2010)
9. Dasgupta, P.: kOthar KriYa kOrmo. Dey's Publication, Kolkata (1987)
10. Emonds, J.E.: A Transformational Approach to English Syntax. Academic Press, NY (1976)
11. Haegeman, L.: Introduction to Government and Binding Theory. Blackwell, Oxford (1994)
12. Haegeman, L., Zanuttini, R.: Negative Heads and the Neg Criterion. The Linguistic Review 8, 233–251 (1991)
13. Harris, A.C., Campbell, L.: Historical syntax in cross-linguistic perspective. Cambridge University Press, Cambridge (1995)
14. Nara, T.: Avahattha and Comparative Vocabulary of New Indo-Aryan Languages. ISCAA, Tokyo (1979)
15. Prasain, B.: A Computational Analysis of Nepali Morphology: A Model for Natural Language processing. Ph.D. Dissertation, Tribhuvan University (2012)
16. Sen, S.: bhaSar Itibritta. Ananda Publishers, Kolkata (2007)

Symmetry in Prosodic Pattern of Rhyme and Daily Speech: *Pragmatics of Perception*

Rimi Ghosh Dastidar

School of Language and Linguistics, Jadavpur University, Kolkata

Abstract. It is out of the debate that prosodic pattern varies according to the contexts. This paper deals with the metrical pattern of the rhyme and explores that this pattern is also subject to change according to the discourses irrespective of a predetermined structure of the verse. And this attempt also helps to understand how the users of a particular language do change their prosodic appearance according to the different situations. This attempt aims to model the strong similarity between the prosodic patterns of the daily speech and rhyme subject to this situational variation.

1 Prologue

Formal linguists usually refer language data which are derived from their own imagination instead of copying the real raw data collected from the bona fide conversation. If a language is to be believed as a body to be examined by the cognitive expertise then the bona fide human conversation would be the best example for this purpose. Real human speech therefore reflected in daily conversation is a total projection of human mind where linguistic and extra-linguistic factors work together. For example a sentence cannot be the part of a real conversation iff there is no impact of pre-association of speaker and hearer in respect to those sentences as referred. The identical sentence reflects deliberate difference in terms of prosodic pattern when it occurs in two different contexts. The speaker exhibits varied prosodic intonation. A sentence or a set of words cannot be comprehensible to the speaker-hearers' domain of conversation if the sentence is not the fact of the bi-directional traffic working in the speaker-hearer's mind and does not contain any prosodic variation.

Keeping the fact of the bi-directional traffic this paper is a deliberate attempt to show how the prosodic (to be more specific 'metrical') pattern of different genres (here normal human conversation and rhyme in specific) can be considered as default symmetry. Now the task is, first to generalize the prosodic pattern in a measurable format and second to identify the texts or speech documents according to the sets which would be provided for data.

2 Prosody in Verse and Prosody in Normal Speech

In case of verse, prosody generally refers to the swiftness of the well-governed and well-measured sound flow. This prosody is a salient feature for rhyme. In a given structure every word has unique meaning irrespective of the rhymes and prosody makes those

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 823–830, 2013.
© Springer International Publishing Switzerland 2013

words move with a flowing 'intention'. In an ordinary literary practice we can measure the differentiated variation of Bangla rhyme and set the structures according to the measurement. This paper claims that if we can measure the daily conversation subject to the different contexts, it can be observed that there are obviously some similarities between the metrical patterns working simultaneously for the daily speech and the rhymes associated with the real contexts. Rhyme is considered as oral literature which is viewed as specific literary genres, i.e. female literature as well as children's literature. Rhymes don't bother to abide by linearity all time and not being framed within a restricted pattern, it flows on its own way and it follows its own undefined nature of flowing. It never allows artificial or mechanical projection. Our regular speech in most of the cases is always generated without this mechanical projection.

Prosodic properties of normal speech are derived from the suprasegmental attributes underlying human utterances. Human speech is not a mere combination of phonemes and grammatical properties. Suprasegmental effects add an implicit melody which works for the dynamic motion of speech. The segmental contents of speech are static and they need prosody as an engine for their movement. Prosody liquefies the words or string of words from their frozenness and binds up the wholeness with equilibrium. Prosody itself is motion. It rises very 'naturally'. Our perception reacts in different way due to variety of external instinct. These different reactions are reflected through different prosodic dimensions. While prosody for verse is deliberately obvious with its metrical pattern, prosodic coding for normal speech moves latently.

Prosodic nature of rhyme normally reflects a tonal trend in daily speech in a given social set up as most of the rhymes reflect kolas of parted pictures of our daily life. Thus, Rhyme/ rhymes is a natural form of human quality that can be used in the spontaneous creation of "naturalistic" art. Rhyme or rhymes are basically traditional oral literature.

3 Metrical Properties of Prosody

In order to show the metrical pattern of verse and normal prosody we here need to be introduced with the three basic metre pattern of verse form. In Bangla, there are three metrical patterns: 1) Syllabic pattern 2) Simple moraic pattern and 3) Composite moraic pattern. Syllable and mora are the basic tools for measuring Bangla verse form. Here we only concentrate on syllable as our primary concern is rhyme and normal prosody and closely syllables mark the statistic measurement for rhyme. Rhymes generally follow syllabic pattern where each syllable is considered as a single unit. For example,

1 1 1 1 1 1 1 1 1 1 1 1

(1) Kho. Ka: gEe.lo/ mach: dhor.te: khir/ no.dir: ku.le ["The boy went for fishing to the bank khir (local name) river"][1]

[1] The principal units of Bengali prosody are verse, clause, foot, sub foot and syllable. These units are determined by different kind of pause; such as full pause, half pause, sub pause, light pause. Full pause section indicates the metrical line of a verse. Half pause section refers to the clause. Light pause section traces the foot.sub foot is regarded as the sub pause section and finally the nuclear pause is considered as syllable. Each foot of verse begins with an accent and ends with pause. In the verse line dot **/.**/ marks the syllable break; **/:/** refers sub foot where slash **///** indicates foot break.

1 1 1 1 1 1 1 1

(2) Chip.khan: tin.daNR/ tin.jon:mal.la ["The long boat, three oars, three boatmen"]

We can adhere in this ground as there is a close symmetrical association between prosody of our daily speech and rhymes/rhyme and we can also scan our daily speech with the syllabic tool. We normally take pause with syllable during speaking. This pause is often implied. If we consider the following example,

(3) 1 1 1 1 1 1 1 1
a.mi to/ ja.bo na/ bol.lam
"I said that I'll not go."

(4) 1 1 1 1 1 1 1 1 1
tui ki ja.niS/ tui ki ko.re.ciS?
"Do you know what you have done?"

(5) 1 1 1 1 1 1 1
O.ma tu.mi aS.be na?
"Alas! Won't you come?"

In case of daily speech the foot break may not be regularly marked and each foot may not carry equal number of syllable where rhymes are decorated with a frequent regular measurable pattern for its versification. Daily speech marks the break according to the context. Speakers draw the pause; strike the stress, show emotion and execute exclamation as they need. If we consider example (4) and explain the statement with different pragmatic context we find different results. /Tui ki janiS tui ki koreciS / can be a consequence of an extraordinary success or achievement of the hearer whom the speaker informs the event with an overwhelming charm and it can be presupposed that the listener is not yet aware of his deed. Again /Tui ki janiS tui ki koreciS / can be said to a person who did a big mistake which would cause a severe mess to him as well as to many. The intonation, prosody and the stress will obviously differ in both case and the varied outburst of the speaker marks foot break at different position for a single sentence.

4 Social Determinism of Prosody

Our basic claim in this attempt is, just to remember the catch line of this paper, to show the similarity between the metrical as well as prosodic ratios of ordinary conversations and verses evolved comparatively from more natural process than poetry. The foremost consideration for this overture is to discuss some social aspects as we are trying to decipher Bangla chORa (Bangla Folk Verse) as a social process. It is well-known in the descriptive sociolinguistic studies that social varieties of speech related to the contexts are best reflected in the spectrum of register.

Now we focus prosody with the perspective of register. Thomas Bertrum Reid introduced a sociolinguistic term 'register' in 1956 which refers to context sensitive language use. An identical concept is explored/ realized through different lexemes/ words depending on its context. A professor in the discipline of Chemistry can normally use the term 'sodium chloride' in his lab to fulfill his purpose but when his domain changes to house, he shifts to salt .This regularly happens in our daily speech.

Register in certain aspect is associated with prosody, that is, according to condition concerned for the scale of language use prosody modifies its physical nature in the range of speech.

(6) GhoriTa kinte kOto khOSali? [*khOSali]* is a very popular Bangla informal term in the meaning for expenditure] ("How much have you spent for this watch?")

Same concept is realized through a completely different sentence.

(7) KOto dam diYe kinle ghori Ta? (How much have you spent for this watch?)

(6) is used in informal domain of speaking as the term 'khOSali' can be used within peer group. (7) is a formal one.

These variations are observed not only in the supplement of words, but also seen in the prosodic variations in the human conversations. Linguistically prosody depends on the articulatory manifestation which involves stress, intonation, pitch and voice quality. These all variable are dependent on physical phenomena of the speaker as well as the concerning words. We consider (8) and (9) in this regard.

(8) Jomunaboti SOrossoti kal jomunar biYe
Jomuna jabe Sosur bari kajitOla diYe
kajiphul tulte giye peYe gelum mala
hat jhumjhum pa jhum jhum Sitaramer khela

(9) Jomunaboti SOrossoti kal jomunar biYe
Jomuna tar baSor rOche buke barud diYe
biSer topr niye।

The first lines of both verses are identical and reflect same prosodic pattern but while we move to second line the theme waves to be changed. (7) reveals an image of marriage ceremony of a girl related to marriage in a pure lightened way and leaves a resonance of happiness while (8) too initiates with a story of would be marriage and denotes weariness and extreme ignited condition of the would be bride. The intrusive effect of the verse generates a depressive prosody. Similar event that is marriage ceremony entails different presupposition. One is totally for child world and the second one belongs to the world of cruelty. So prosody is also highly determined by the contextual appropriateness and the semantic weight of the words. We will show this difference this difference through a technical way to prove its liability. Consider fig a.1 and a.2.

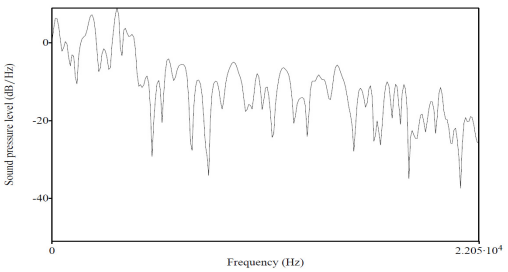


Fig. 1. a.1: Prosody of rhyme in charm

In Figure a.1, in the diagram horizontal position represents frequency and vertical position represents amplitude. The total bandwidth is 22050.00 hertz. Highest amplitude is 9.0 db where lowest is -51.0 dB

Bandwidth= the difference between high and low frequency.
Amplitude= measure of change of periodic variable over a single period.

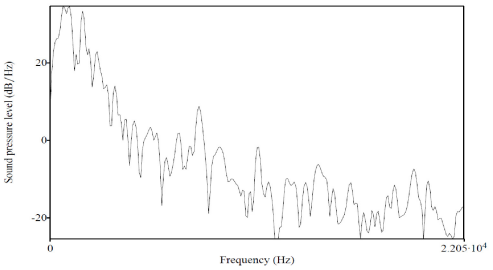


Fig. 2. a.2: Prosody of rhyme in depression

Fig a.2 shows the diagram where total bandwidth is 22050.00 but amplitude varies. High is 34.6 dB and low amplitude is -25.4 dB.[horizontal and vertical position indicate same as fig a.1]

We here want to show the variation in graph for the two structurally related rhymes. At a same frequency level the amplitude differs with the prosodic variation. Extra linguistic factors correlate with linguistic contents.

We examined the alternation of prosody not only in case of rhymes, but our survey provides evidence in regular speech too. An identical speech has been found against a same question to some speaker without making them aware about our intention and it clearly marks the difference. Variations are observed not only in the supplement of words, but also seen in the prosodic variations in the human conversations given sentence was (10)

(10) Q: kEmon colche apnader? [how are you doing?]

A: amader abar jibon! e jiboner na ache din ; na ache rat. [life of us! There is neither day, nor night]

The physical contexts were separate. One context was in a village where few days ago terrible storm caused a toll of death and utterly messed up everything and survivors are spending sleepless day. And the second context belongs to a corporate office where the speakers have to engage 24x7 hours work with a lump sum amount of salary. Identical Conversation with identical meaning explores different prosody. Fig (a.3) and Fig (a.4) will show the variation deliberately.

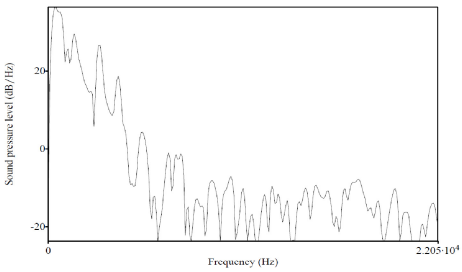


Fig. 3. (a.3): Prosody of Regular Speech in Hopelessness

Fig a.3 shows the diagram where total bandwidth is 22050.00 .High is 36.3 dB and low amplitude is -23.7 dB.[horizontal and vertical position indicate same as fig a.1.

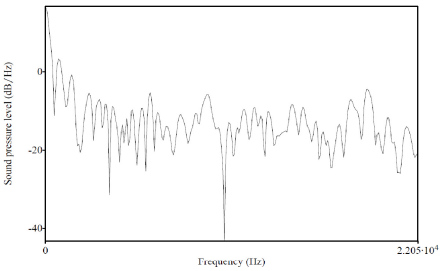


Fig. 4. (a.4): Prosody of Regular Speech in twitchy mood

Fig a.3 shows the diagram where total bandwidth is 22050.00 .High is 16.8 dB and low amplitude is -43.7 dB.[horizontal and vertical position indicate same as fig a.1.

If we compare fig a.3 with fig a.4 we have a clear difference. The statements are identical in configuration though they vary in intonation. And the variance is obvious from the difference in amplitude.

If we consider fig a.1 and fig a.4 as these two diagrams show the spectrum graph of such verse and statement which imply a kind of energetic mood and restlessness. And surprisingly the amplitude for two cases correlates (see table 1) Amplitudes are approximately closer to each other in respect of high and low level 9, 16 and -51, -43 in two figures.

And if we compare fig a.2 and fig a.3 these two diagrams display spectrum graph of those verse and statement which entail a depressive mood with a heart broken extreme pain. Let's see the amplitude in table 1.

Here too the amplitudes draw a parallel measurement with each other in 34, 36 and -25,-23 in high and low levels respectively.

Table 1. Amplitude showing Correlation between Tonal Variation in daily speech and rhyme

Figures	Amp (high) (dB)	Amp (Low) (dB)
a.1 (Prosody of rhyme in charm)	9.0	-51.0
a.2 (Prosody of rhyme in depression)	34.6	-25.4
a.3 (Prosody of Regular Speech in Hopelessness)	36.3	-23.7
a.4 (Prosody of Regular Speech in twitchy mood)	16.8	-43.2

It is really amazing to find out such an association between rhyme and daily speech in terms of prosody where different related exclamation irrespective of rhyme and daily speech show an alliance in spectrum.

Besides spontaneous speech, reading texts too result in a similar way. The words of the text impose a cognitive introspection on the reader to differentiate in prosody. We consider (11a) and (11b) to prove this.

(11a) SOkal theke briSti hocche. Khokoner aj bhari mOja.Iskul jete hObe na. tar opor maYer kache sunece dupure hObe khicuri ar iliS mach bhaja. Ilish mach to bhison priyo khokoner. Barir samne besh anekta jol jomece.Ekta nouka bhaSiYe dile kEmon hoy? jEmon bhaba tEmon kaj. Khokon chutte giye porar ghor theke kagoj ene bose porlo nouka banate. Ah! Erokom din Jodi roj aSto khokoner kache? [*It's raining from morning. Khokon is too happy today. He can be off from school today. Again he heard about lunch menu which will have khichri (hotchpotch) and Hilsa fry. Hilsa fry is too favourite to Khokon . The front side of their house is already turned to be a stream. How is that if a paper boat will be floated? As soon as he thinks, he rushes to study room, brings paper and sits to form it a boat. Ah! If such day comes regular to khokon!*]

(11b) SOkal theke briSti hocche. Caridike kEmon Ekta Ondhokar gumot bhab.khokoner ar bhaloi lagce na.Saradin ei bhabe barite boSe thaka. Bondhuder sathe dEkha hObe na aj. O khelbe ta kar SOnge? ma to nijer kajei bEsto.bapio laptope much guje pore ache.Ektu golpo korar o nei keu. cupcap nijer ghOre ese boSlo khokon. [*It's raining from morning. The surroundings show a drowsy, gloomy and dull weather. Khokon feels too boring. He has to be in home whole day. He can't meet his friends. Who will be his co-player? Mom is busy with her own work, laptop arrests dad. There's nobody to talk. He comes to his room and sits quietly*]

These two different texts are realized in different prosodic pattern. Both of them are oriented with a child's behavior in a rainy day. But the readers' approach and voice quality varies with the plots. Prosody, a naturally born and free flowing feature of utterance, can never ignore discourse in its appearance. Human beings communicate prosodic units behind which the contextual behavior strongly works.

5 Conclusion

Our prior concern is to reach to the inference rhymes and daily speech share an approximate close ratio in terms of prosody. Rhymes are the masonry of pictures from daily life. They are too common and too inseparable from human life and their linguistic structure is extraordinarily inherent. Rhymes share a surprising attribute in

the sense that they don't have any birth recognition, that is, nobody knows when they were originated and who created them. So we can suppose them as very natural rather than artificial or mechanical and supposedly universal rather than universal and these are distributed diachronically among the generation. Daily speech born of innateness manipulates a prosodic balance over its performance. Human beings too articulate both them with an easy motion without facing any obstruction in the performance level of speaking.

This may be the right time to forecast that there are some possibilities of this claim made in this paper subject to the applicability. It is very easy to say that if this symmetry between rhymes and ordinary conversation measured in the main course of this paper has been evaluated and proved, this will offer a systematic protocol for the universal tendency of natural prosody. It may help in future to evaluate human speech in an extraordinary manner for several cognitive computational researches.

References

1. Dasgupta, P.: The Centrality of the Mora in Bangla Prosody. Paper presented in National Seminar on Prosody, at University of Calcutta,organized by the Department of Linguistics, University of Calcutta & CIIL, Mysore (2002)
2. Dasgupta, P.: Osthir chOnder chandoggo. Gangeya Patra 19, 26–35 (2003)
3. Ray, P.S., Hai, M.A., Ray, L.: Bengali Language Hand book. Center for Applied Linguistics, Washington D.C (1966)
4. Ghosh, S.: SONkho ghoSer kobita SONgroho.kolkata. Dey'z publishing (2009)
5. Roy, A.: chOra SOmogro. BaniSilpo, Kolkata (2010)
6. Tagore, R.: chOndo. In: Collected Works of Rabindranath, vol. 10, Govt. of WB Publication, Kolkata (1989)
7. Tagore, R.: lokSahitto. In: Collected Works of Rabindranath, vol. 10, Govt. of WB Publication, Kolkata (1989)
8. Chomsky, N., Halle, M.: The Sound Pattern of English. Harper & Row, New York (1968)
9. Hogg, R., McCully, C.: Metrical Phonology. Cambridge University Press, Cambridge (1987)

Prosody Modeling: A Review Report on Indian Language

Sudipta Acharya and Shyamal Kr. Das Mandal

Indian Institute of Technology Kharagpur, India
sudipta.acharya2009@gmail.com

Abstract. This paper presents a detail study on prosody parameters such as Pause, Duration, F_0 and Intensity, and different methods for their modeling for Indian language. Various Speech Synthesis Systems are now appearing for some of the major Indian languages; however, all of these can only generate flat and monotonous speech – raising perceptual difficulties to sustain listening. Prosody (intonation and rhythm) of spoken language plays an important role for intelligibility and naturalness in synthesized speech.

Keywords: Text-to-Speech Synthesis (TTS), prosody, F_0 modeling, pause modeling, duration modeling.

1 Introduction

Indian languages lack of speech applications. Speech is the best medium of communication, so it's obvious for the people to expect to be able to communicate with computers through speech. With about 60% people are illiterate in Indian subcontinent, so there is a huge socio-economic importance of applications involving conversion of text into speech, which is commonly known as text-to-speech (TTS) system. TTS system has a great application in Indian language – such a system is useful to overcome the literacy barrier of the common masses, empowering the visually impaired population, increasing the possibilities of improved man-machine interaction through online newspaper reading from the internet and enhancing other information systems.

2 Objective

A review of different methods for prosody modeling of parameters such as Pause, Duration, F_0 and Intensity for Indian languages.

3 What Is Prosody?

'Prosody' is defined as a supra segmental feature of speech, which determines the pitch, duration (quantity), and loudness (quality) in single speech sounds and accent and rhythm in sequences of sounds. This can be interpreted in a way that prosody denotes certain features of speech, whose measurable correlates can be found in the

R. Prasath and T. Kathirvalavakumar (Eds.): MIKE 2013, LNAI 8284, pp. 831–842, 2013.
© Springer International Publishing Switzerland 2013

fundamental frequency, segment duration and intensity[14]. Furthermore, it can be derived that prosodic features occur on the phoneme level as well as on the level of syllables, words and phrases. Fujisaki proposes that "prosody is systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production." Its realization involves both segmental and supra-segmental features of speech, and is influenced, not only by linguistic information, but also by para-linguistic and non-linguistic information[11][13]. It can be represented either explicitly by the written language, or can be easily and uniquely inferred from context [33]. Para-linguistic information is defined as the information that is not infer able from the written counterpart, but is deliberately added by the speaker to modify or supplement the linguistic information. A written sentence can be uttered in various ways to express different intentions, attitudes, and speaking styles, which are under the conscious control of the speaker. For example, the information regarding whether a speaker's intention is an assertion or a question is discreet, but it can also be continuous in the sense that a speaker can express the degree within each category. Non-linguistic information: It concerns such factors as the age, gender, idiosyncrasy, physical and emotional states of the speaker, etc. These factors are not directly related to the linguistic and Para-linguistic contents of the utterances and cannot generally be controlled by the speaker, though it is possible for a speaker to control the way of speaking to intentionally convey an emotion, or to simulate an emotion, as is done by actors.

4 Methods Already Used

4.1 Pause Modeling

Rule Based Method: It needs linguistic expertise and a large corpora to set up the rules. Advantages are, this is a very simple and convenient approach. But time-consuming to get many rules.

ANN Model: It gives better results than Rule based approach. It needs a very large corpus for training, but the results are still unsatisfactory.

Statistical Linear Approach: This method is proposed by Fujisaki et. al. 1989. For occurrence and duration of pause the model equations will be in linear form [8][9].

$$Pq=\min\{\max(a*l + b*d + c, 0),1\} \quad (1)$$

$$Dq= \alpha*l+\beta*d+\gamma \quad (2)$$

where l is the length of an uninterrupted phrase is to be expressed by the mora/syllable. d is the distance between the current phrase and its dependant counterpart. P is the probability for pause occurrence, its value should be within (0,1). q is the phrase type. Coefficients a, b, c can be determined by minimizing the mean squared error over the ranges of l and d specified in equation(1). Equation(2) describes the model equation for pause duration, where α, β, γ are the coefficients and determined by linear regression analysis by using the training data.

4.2 Duration Modeling

In duration modeling duration can be modeled in two ways. One is segment duration or phoneme duration, another one is syllable duration. Again it can be divided in two categories:

Rule Based Models

i)Klatt duration model: Segmental duration can be formulated as

$$D_d=(D_a-D_b)*P/100 +D_b \quad (3)$$

where ***Da*** is the inherent phone duration in ms., ***Db*** is minimum phone duration and ***P*** is the factor determined by applying Klatt's eleven rules[48]. For instance, one rule may be, phoneme duration in case of a vowel or a syllabic constant occurring after a pause can be determined by the formula which assumes ***P*** should be 140 for this particular case [34].

ii)IITM duration model: A set of 31 rules comprising of IF-THEN structure were derived from an analysis of 500 sentences spoken by a native male speaker while considering the positional and contextual factors[48].

Statistical models

i)Sum-of-Products (SOP)model. ii)Classification and regression Trees(CART) model.

iii)Neural Network(NN) model.

4.3 F_0 Modeling

Phonological

i) Tones and Breaks Indices (ToBI)

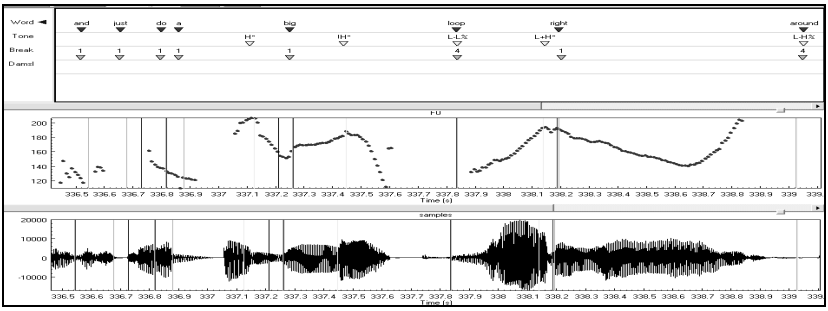


Fig. 1. An example of ToBI transcription[15]

ToBI is the most widely used system for the symbolic transcription of intonation at present. It provides a four level transcription system. Six different pitch accents (H*, L*, L+H*, L*+H, H+L*, H*+L) and two levels of intonational phrasing (intermediate and full intonational phrase). Pitch accents are mainly aligned with accented syllables. A boundary tone is associated with each intonational phrase boundary. The symbol L- (H-) describes a low (high) tone at an intermediate phrase boundary. The symbols

L-L%, L-H%, H-L% and H-H% are used to represent full intonational phrase boundaries.

Break indices mean boundaries between words and come in five levels[41].

0 : Clitic boundary, **1**: normal word-word boundary. **2** : Apparent intonational boundary. **3:** Intermediate phrase boundary. L- or H-. **4:** Intonation phrase - phrase or sentence final L or H, marked with L% or H%.

ii)IPO.

Acoustic Phonetic

i) Fujisaki Superposition model

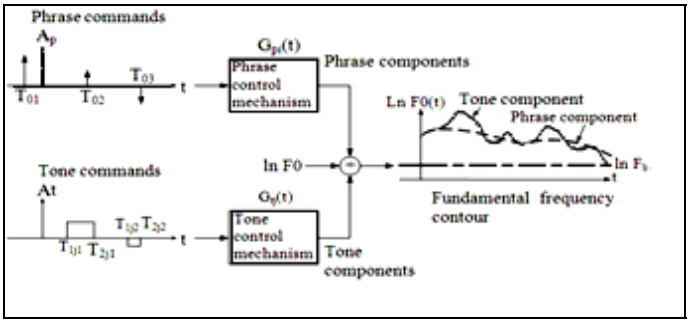


Fig. 2. Fujisaki model for generation of F_0 contour [12]

Seven derived output parameters are computed from the conventional parameters. It is noticed that these derived parameters mostly reflect the geometrical appearance of the F_0 contour of the speech. They are baseline frequency F_b, number of phrase commands Ap, which are impulse response, number of tone commands At, which are stepwise functions, phrase command duration, tone command duration, amplitude of phrase command and amplitude of tone command. The logarithmic F_0 contour is the superposition of logarithmic base line frequency, Accent command and phrase command.

Other Models

i)Tilt model. ii)Neural Network model. iii)IITM intonation model.

5 Literature Review

5.1 Pause Modeling

It's very important to predict the prosody information for producing natural sounding speeches. The pause duration model, which is one of the important parts of prosody model, is essential in improving prosodic quality. The synthesized speech will not be natural and even unacceptable sometimes, if we only use constant pause duration for each prosody boundary. Although the research on pause model has not been paid as much attention as pitch model and duration model before, there is still some work have been done in the last several years[59]. For an example, the sentence “The

daughter of the man and the woman arrived" could either mean that the daughter and the woman arrived or that only the daughter arrived. Speakers can prevent such syntactic ambiguities, in principle, by varying prosody—an exaggerated pause after "man" to convey one meaning, a pause after "woman" to convey the other[32]. Indeed, such prosodic cues like pause can be used to disambiguate many utterances. For an utterance, pause is also very important to get a clear idea about prosodic phrase or clause.

The rule-based method[28] for is one of the typical methods, which used linguistic expertise to infer some pause generation rules based on observations on large corpus. This approach is simple and convenient, but it's quite time-consuming to get many trivial rules, and the results, which influence the prosody generation, were not so good. Later on, Chen Sin-Horng et. al. 1992 tried to use the training model such as ANN for pause duration. It generated better results than traditional rule based methods, but with ANN model, we had to prepare a very large corpus for training, and the results were still limited in most cases. It is also hard to get the relationship between pause duration and other features with only ANN outputs. Also, the results are still unsatisfactory in most cases. Similar work has also been done by some others. Fujisaki et al.[9] developed linear models for predicting the occurrence probability and duration of pauses in readout speech of Common Japanese [8][9]. The works on pause model were also studied for Chinese, Mandarin, Japanese and Bengali [5].

Review on Indian Language

Das Mandal *et. al.*[5] developed a statistical linear model for predicting the occurrence probability and duration of pause for a readout speech. This paper reported the results of a detailed study on the various factors (i.e. syllable length and depth) that have influences on the occurrence of sentence-medial pauses and their duration for Bangla read-out text. The reason for this study is to develop models to pause insertion and pause length assignment in a read-out mode Bangla text-to-speech synthesis system. The occurrence and duration of sentence-medial pause after a verb phrase is higher than the pause after any other type of phrase.

5.2 Duration Modeling

Duration is one of the most important prosodic feature that contributes to the perceived naturalness of synthetic speech. Variation in segmental duration serves as a cue to the identity of a speech sound and helps to segment a continuous flow of sounds into words and phrases thereby increasing the naturalness and intelligibility. In natural speech, segmental durations are highly context dependent. For an instance, suppose a vowel /e/ that is as short as 35ms. in the word "pehlA" and as long as 150 ms. in the word "rahegA". So the primary goal in duration modeling is to model the natural speech duration pattern, by taking various features which has impact on the pattern. An important restriction is that because of the nature of the Text-to-Speech synthesis problem, only those features that can be automatically derived from text can be considered[24][26]. Duration plays as much important role as intonation in the encoding/decoding of speech by the speaker/listener. The duration is a part of the prosody and contain important cues for understanding the spoken text. Variations in duration provide assistance for the listener to understand the meaning. The duration

models can be split into two as rule-based and statistical duration models [34][43]. Rule based methods does not work for large amount of data. This method depends on linguistic and phonetic literature about the factors that affect duration of the units (segments, syllables or phones). In general, rule based methods are difficult to study, due to complex interaction among the linguistic features at various levels. Statistical data-driven methods are attractive when compared to rule based methods. This method works when large phonetically rich sentences are present in the corpora.

Duration can be defined as the time taken to utter an acoustic unit, i.e. phoneme, syllable. Duration modeling studies mainly concentrate on phoneme duration [23][24][30][31][29][37]. However, there are studies also on syllable duration [2][3][28][48] [50].

As a rule-based approach, Klatt used the notion of intrinsic duration introduced by Peterson and Lehiste (1960) [2][23]. Intrinsic duration is the average duration of the syllable nucleus. His model assumes that each phonetic segment type has an inherent duration that can be modified by a set of eleven rules, but phonemes cannot be compressed shorter than a certain minimum duration[23]. Similar models were developed for other languages like French and German.

Riley (1990, 1992) used a 1500 hand-labeled speech database from a single male speaker for segmental duration prediction using CART (Classification And Regression Tree) [2][37]. van Santen(1994) states that classification trees require a huge amount of training data to cover all possible feature space and proposed the Sum-of-Products models reference. CART models were developed for the languages Czech, Turkish[39], Korean[4], Hindi and Telegu[24]. CART approach was used in this research to model phone duration of Lithuanian using very large speech corpus[38]. 300 thousand samples of vowels and 400 thousand samples of consonants extracted from VDU-AB20 corpus were used in experimental part of research. The achieved results, correlation for vowels is 0.8 and for consonants is 0.75, and RMSE of ~18 ms. are comparable with those reported for Check, Hindi and Telugu, Korean. Sum-of-Product model finds phoneme durations by a summation of functions of attributes[37]. Campbell (1992) used Neural Networks for predicting syllable timing[2]. He used a categorical factor analysis to find out the factors that influence the syllable duration. Other neural network based models were also developed for Spanish, Arabic, German and Portuguese and this knowledge was taken from Rao, K.S. 2012[46].

Review on Indian Language

Kumar S.R.R. *et. al*. (1989,1990) used Rule based method for Hindi Language[27]. This is a syllable based duration modeling. This speech synthesis system was based on the parameter concatenation (Linear Predictive). In order to improve the quality of speech produced from input text, it is necessary to expand the set of rules. Another Rule based method for syllable duration modeling is used by Roy R. *et. al.*[41][49] for Bengali language. But this method does not consider the effect of some linguistic factors like POS of the word and the type of phrase or clause. Rao K.S *et. al*.[43] also used syllable duration modeling by using four layered Feed Forward Neural Network (FFNN) for Hindi, Telugu and Tamil. 88% of the syllable durations are predicted within 25% of the actual duration by this method. Performance can be further

improved by including the accent and prominence of the syllable in the feature vector and accuracy in labeling, diversity of data in the database, and fine tuning of neural network parameters. Krishna N. S. *et. al.*[24] built up a model for phoneme duration by using CART method for only Hindi language. Again, in 2004 N.S.Krishna *et. al.*[24] developed a phoneme duration model for Hindi and Telegu languages. The work was not in large annotated speech corpora, but for better prosody learning a large corpus is needed. The limitations are firstly, Corpora used here is not optimal and data sparsity problem is not taken care of, and secondly, modeling and analysis is done on smaller data set. Rao K.S *et. al.*[46] also worked on syllable duration for Hindi, Telugu and Tamil by the use of support vector machine (SVM) and did the comparison with FFNN method. 86% of the syllable durations are predicted within 25% of the actual duration. The models are evaluated by computing the standard deviation, correlation coefficient and mean absolute error between predicted and actual syllable. Gopinath D.P. *et. al.*[20] used a statistical method for phoneme duration model for Malayalam language. In this method the duration patterns of vowels and consonants are compared.

5.3 F_0 Modeling

For the generation of highly natural synthetic speech, proper control of F_0 is of primary importance. F_0 is a highly variable parameter determined by both the physical and the linguistic aspects of speech production. When we speak we do not only produce a sequence of speech sounds, but also impose stress and intonation patterns to convey a meaning. General assumption for intonation modelling is that it can be successfully handled only by fundamental frequency, thus, the goal is to develop a model to generate fundamental frequency contours. F_0 is a highly variable acoustical parameter. It is determined by both the physical and the linguistic aspects of speech production. While the value of F_0 measured at a particular instant indicates the vibration frequency of the speaker's vocal cords, the time varying F_0 contour carries abundant information about the linguistic structure of the sentence. The linguistic significance of F_0 is language-dependent[33]. Various intonation models have been proposed in the past. They are contrasted by different viewpoints: the systems may be phonological or phonetic; pitch contours can be produced by parametric or nonparametric methods; or the systems may use level tones or pitch movements. Among the phonological models the most influential one is Pierrehumbert's model, which later evolved into a standard (Tones and Breaks Indices, ToBI) for transcribing American English. As stated in Silverman et. al. (1992), ToBI is the most widely used system for the symbolic transcription of intonation at present. It provides a four level transcription system to the researchers, which obeys the general outline proposed by Beckman and Pierrehumbert[40]. Beckman and Pierrehumbert proposed six different pitch accents (H*, L*, L+H*, L*+H, H+L*, H*+L) and two levels of intonational phrasing (intermediate and full intonational phrase)[40][51]. Pitch accents are mainly aligned with accented syllables. A boundary tone is associated with each intonational phrase boundary. The symbol L- (H-) describes a low (high) tone at an intermediate phrase boundary. The symbols L-L%, L-H%, H-L% and H-H% are used to represent full intonational phrase boundaries. This model was implemented for the languages German, English,

Chinese, Navajo and Japanese [45]. Another example of phonological models is the Instituut voor Perceptie Onderzoek's (IPO) perceptual model which relies on identifying perceptually relevant pitch movements and approximating them with straight lines. The main point of the approach is to simplify the F_0 curve and preserve the same melodic impression to the listener. The IPO model was first generated for Dutch, later it was used for English, German and Russian.

Parametric models that belong to the broader class of phonetic models use a set of continuous parameters to describe intonation patterns. A well-known parametric model is the Fujisaki's Generation process model. The actual F_0 contour is obtained by the superposition of baseline F_0, phrase and accent components on a logarithmic scale. The response of a second order critically damped linear filter is called phrase command, which helps generate the global variation or phrase component of a F_0 contour. Accent component is generated by another second-order, critically damped linear filter in response to a step function called accent command[12][13][36][34]. The model has since been extended to apply to a number of languages including Japanese, Cantonese, Bangla[6], Basque, American English, Chinese, German[34], Korean[10], Spanish and Swedish[35][16]. It was argued, however, that the model would fail to generate some F_0 contour sections commonly found in utterances of British English [7].

The Tilt intonation modeling proposed by Taylor can be considered both as phonological and phonetic because continuous tilt parameters are computed only at event locations and non-event parts of the pitch contour are generated by linear interpolation. Pitch accents and boundary tones are defined as events. Events have rise-fall patterns. Each event is represented by three tilt parameters: duration, amplitude and tilt[53][55]. Duration is the sum of the rise and fall durations. Amplitude is the sum of the magnitudes of the rise and fall amplitudes. The tilt parameter is a dimensionless number which expresses the overall shape of the event[54]. INTSINT (INternational Transcription System for INTonation) – proposed by Hirst and Di Cristo[20]. It is an intonation transcription system which codifies F_0 patterns using a set of abstract tone symbols. Those symbols can be absolute or relative symbols. The {T, M B} symbols, (Top, Mid, Bottom), are absolute symbols for the F_0 variance range of a speaker. The {H, S, L, U, D} symbols, (Higher, Same, Lower, Up stepped, Down stepped), are relative to the previous target-point [11]. Nonparametric approaches use F_0 values themselves. Samples from the pitch contour are taken to develop intonation models. Examples of nonparametric methods are rare. Black and Hunt (1996) used a Linear Regression Based method to predict F_0 target values for the start, mid-vowel, and end of every syllable[1]. Several models based on Neural Network principles are described in the literature for predicting the intonation patterns of syllables in continuous speech[50][21]. Scordis and Gowdy used Neural Networks in parallel and distributed manner to predict the average F_0 value for each phoneme, and the temporal variations of F_0 within a phoneme [50].

Review on Indian Language

An intonation model for the Indian language Hindi was proposed by Kumar A.S.M. *et.al.[*25]. It was shown that intelligibility and naturalness of the synthesized speech improved significantly after incorporation of the intonation rules. They also described some features of the fundamental frequency contours of speech for declarative sentences as well as for interrogative sentences and activated these to an unrestricted

Text-to-Speech system for Hindi. Rao K. S. *et al.* [43] used the Neural Network method for F_0 modeling for the languages Hindi, Telugu and Tamil. The paper shows 88% of the F_0 values (pitch) of the syllables could be predicted from the models within 15% of the actual F_0. The performance of the intonation models is evaluated using objective measures such as average prediction error μ, correlation coefficient γ and standard deviation δ. Listening test was also used to predict the accuracy of the intonation models. The proposed prediction performance of the intonation models using neural networks is compared with Classification and Regression Tree (CART) models. Das Mandal *et. al.*[6] implemented Fujisaki's superposition model for Bangla intonation modeling. Reddy V.R. et.al.[46] proposed an intonation model using feed forward neural network (FFNN) for syllable based text to speech (TTS) synthesis system for an Indian language Bengali. The proposed intonation model predicts initial, middle and final positions of F_0 values of each syllable. The prediction performance of the neural network model is compared with the Classification and Regression Tree (CART) model.

5.4 Intensity Modelling

It has often been discussed in (Rompertl J. et. al. 2007) literature that intensity (or loudness – as a psychological correlate of intensity) is of far less importance than fundamental frequency with respect to supra segmental features of speech. Therefore, our prosody model pays significantly less attention to it[47]. Moreover, we have undertaken theoretical considerations of modeling intensity analogically to fundamental frequency. However, since intensity is much more connected with segmental qualities of speech, the application of such a model is not as straightforward as in the case of fundamental frequency (intensity can be treated as distinguishing feature of a phoneme, unlike F_0 that is basically present at voiced phonemes and not present at unvoiced phonemes).

References

1. Black, W.A., Hunt, J.A.: Generating F0 Contours From ToBI Labels Using Linear Regression. In: Proceedings of Third International Conference on Spoken Language Processing, ICSLP 1996, Philadelphia, USA (1996)
2. Campbell, N.: Timing in Speech: A Multi-Level Process. In: Horne, M. (ed.) Prosody: Theory and Experiment, Kluwer Academic Publishers, Dordrecht (2000)
3. Chen, S.H., Lai, W.H., Wang, Y.R.: A New Duration Modeling Approach for Mandarin Speech. IEEE Transactions on Speech and Audio Processing 11(4) (2003)
4. Chung, H.: Segment duration in spoken Korean. In: Proc. Int. Conf. Spoken Language Processing, Denver, Colorado, USA, pp. 1105–1108 (September 2002)
5. Das Mandal, S., Saha, A., Basu, T., Hirose, K., Fujisaki, H.: Modeling of Sentence-medial Pauses in Bangla Readout Speech: Occurrence and Duration, Interspeech 2010, Makuhari, Japan, September 26-30 (2010)
6. Das Mandal, S., Warsi, H.A., Basu, T., Hisore, K., Fujisaki, H.: Analysis and Synthesis of F_0 contours for Bangla readout speech. In: OCOCOSDA 2010 (2010)

7. Fujisaki, H., Ohno, S., Yagi, T., Ono, T.: Analysis and interpretation of fundamental frequency contours of British English In terms of a command-response model. In: ICSLP 1998 (1998a)
8. Fujisaki, H., Ohno, S., Yamada, S.: Analysis Of Occurrence Of Pauses And Their Durations In Japanese Text Reading. In: ICSLP 1998 (1998)
9. Fujisaki, H., Ohno, S., Yamada, S.: Factors Affecting the Occurrence and Duration of Sentence-medial Pauses in Japanese Text Reading. In: Proc. ICPhS 1999, San Francisco, vol. 1, pp. 659–662 (1999)
10. Fujisaki, H.: Analysis and modeling of fundamental frequency contours of Korean utterances — A preliminary study —. In: Lee, H.B. (ed.) Phonetics and Linguistics — in honor of Prof., pp. 640–657 (1996)
11. Fujisaki, H.: Information, Prosody, and Modeling — with Emphasis on Tonal Features of Speech (Plenary Keynote Paper). In: Proceedings of Speech Prosody 2004, Nara, Japan, pp. 1–10 (2004)
12. Fujisaki, H., Hirose, K.: Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese. Journal of the Acoustical Society of Japan 5, 233–242 (1984)
13. Fujisaki, H.: Prosody, Information, and Modeling - with Emphasis on Tonal Features of Speech –. In: Proceedings of Workshop on Spoken Language Processing, Mumbai, India (Invited Keynote Paper) (2003)
14. Fujisaki, H.: Prosody, Models, and Spontaneous Speech. In: Sagisaka, Y., Campbell, N., Higuchi, N. (eds.) Computing Prosody, pp. 27–42. Springer, New York (1996)
15. http://www.cs.indiana.edu/~port/teach/306/tobi.summary.html
16. Fujisaki, H., Ljungqvist, M., Murata, H.: Analysis and modeling of word accent and sentence intonation in Swedish. In: Proceedings of 1993 International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 211–214 (1993)
17. Fujisaki, H., Ohno, S., Yamada, S.: Analysis of occurrence of pauses and their durations in Japanese text reading. In: ICSLP 1998 (1998)
18. Fujisaki, H., Ohno, S., Yamada, S.: “Factors Affecting the Occurrence and Duration of Sentence-medial Pauses in Japanese Text Reading. In: Proc. ICPhS 1999, San Francisco, vol. 1, pp. 659–662 (1999)
19. Gopinath Deepa, P., Vinod Chandra, S.S., Veena, S.G.: A hybrid duration model using CART and HMM. In: Proceedings of IEEE, TENCON 2008 (2008)
20. Hirst, D.: Automatic Analysis of Prosody for Multi-lingual Speech Corpora. In: Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (eds.) Improvements in Speech Synthesis, Cost 258: The naturalness of synthetic speech, pp. 320–327. John Wiley & Sons, West Sussex (2002)
21. Hwang, S.H., Chen, S.H.: Neural-network-based F0 text-to-speech synthesizer for Mandarin. IEE Proc. Image Signal Processing 141, 384–390 (1994)
22. Klatt, H.D.: Review of Text-to-Speech Conversion for English. Journal of the Acoustical Society of America 82, 737–793 (1987)
23. Krishna, N.S., Murthy, H.: Duration modeling of Indian languages Hindi and Telugu. In: 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA, pp. 197–202 (May 2004)
24. Krishna, N.S., Talukdar, P.P., Bali, K., Ramakrishnan, A.G.: Duration Modeling for Hindi Text-to-Speech Synthesis System. In: Proceedings of International Conference on Spoken Language Processing, ICSLP 2004, Korea (2004)
25. Kumar, A.S.M., Rajendran, S., Yegnanarayana, B.: Intonation component of text-to speech system for Hindi. Computer Speech and Language 7, 283–301 (1993)

26. Kumar, S.R.R., Yegnanarayana, B.: Significance of durational knowledge for speech synthesis in Indian languages. In: Proc. IEEE Region 10 Conf. Convergent Technologies for the Asia-Pacific, Bombay, India, pp. 486–489 (November 1989)
27. Lin-Shan, L., Chiu-Yu, T., Ming, O.-Y.: The Synthesis Rules in a Chinese Textto-Speech System. IEEE Trans. Acoustic, Speech, Signal processing 37(9), 269–285 (1989)
28. Lee, S., Oh, Y.W.: Tree-Based Modeling of Intonation. Computer Speech and Language 15, 75–98 (2001)
29. Lee, S., Oh, Y.W.: Tree-Based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems. Speech Communication 28, 283–300 (1999a)
30. Lee, S., Oh, Y.W.: CART-Based Modeling of Korean Segmental Duration. In: Proceedings of Oriental Cocosda Workshop (1999b)
31. Lehiste, I., Olive, J.P., Streeter, L.A.: Role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America 60, 1199–1202 (1976)
32. Li, Y., Lee, T., Qian, Y.: " Analysis and Modeling of F0 Contours for Cantonese Text-to-Speech" TALIP. TALIP 3(3), 169–180 (2004)
33. Mixdorff, H.: An integrated approach to modeling German prosody PhD thesis, Technical University, Dresden, Germany (July 2002)
34. Mixdorff, H., Fujisaki, H.: Analysis of voice fundamental frequency contours of German utterances using a quantitative model. In: Proceedings of 1994 International Conference on Spoken Language Processing, vol. 4, pp. 2231–2234 (1994)
35. Mixdorff, H.: A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. In: Proceedings of ICASSP 2000, Istanbul, Turkey, vol. 3, pp. 1281–1284 (2000)
36. Möbius, B., van Santen, J.P.H.: Modeling Segmental Duration in German Text-to-Speech Synthesis. In: Proceedings of International Conference on Spoken Language Processing, ICSLP 1996, Philadelphia, USA, October 3-6, vol. 4, pp. 2395–2398 (1996)
37. Norkevičius, G., et al.: Modeling Phone Duration of Lithuanian by Classification and Regression Trees, using Very Large Speech Corpus. Informatica 19(2), 271–284 (2008)
38. Öztürk, Ö., Çiloğlu, T.: Segmental duration modelling in turkish. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 669–676. Springer, Heidelberg (2006)
39. Pierrehumbert, J.: Tonal Elements and Their Alignment. In: Horne, M. (ed.) Prosody: Theory and Experiment, Kluwer Academic Publishers, Dordrecht (2000)
40. Roy, R., Basu, T., Saha, A., Basu, J., Mandal, S.D.: Duration modeling for Bangla text to speech synthesis system. In: International conference on Asian Language Processing, Thailand (2008)
41. Rajeswari, K.C., Maheswari, P.U.: Prosody Modeling Techniques for Text-to-Speech Synthesis Systems: A Survey. International Journal of Computer Applications 39(16) (2012)
42. Rao, K.S., Yegnanarayana, B.: Modeling Syllable Duration in Indian Languages Using Neural Networks. In: Proceedings Int. Conf. Acoust. Speech Signal Processing, Montreal, Quebec, Canada, pp. 313–316 (2004)
43. Rao, K.S., Yegnanarayana, B.: Intonation modeling for Indian languages. Computer Speech and Language 23, 240–256 (2009)
44. Rao, K.S.: Predicting Prosody from Text for Text-to-Speech Synthesis. Springer Briefs in Electrical and Computer Engineering. Springer Science Business Media, New York (2012), doi:10.1007/978-1-4614-1338-7
45. Sreenivasa, R.K., Yegnanarayana, B.: Modeling syllable duration in Indian languages using support vector machines. In: Proc. 2nd Int. Conf. Intelligent Sensing and Information Processing,ICISIP-2005, Chennai, India (January 2005)

46. Reddy, V.R., Rao, K.S.: Intonation modeling using FFNN for syllable based Bengali text to speech synthesis. In: Computer and Communication Technology(ICCCT), pp. 334–339 (2011)
47. Romportl, J., Kala, J.: Prosody modelling in Czech Text-to-Speech synthesis. In: The Proceedings of Sixth International Workshop on Speech Synthesis (2007)
48. Roy, C., Basu, T., Saha, A., Das Mandal, S.K., Datta, A.K.: Studies on Duration of Steady States and Transitions in V-V Combination in Bangla Words. In: Proc. of FRSM-2008, Kolkata, India, pp. 157–160 (2008)
49. Roy, R., Basu, T., Basu, J., Saha, A.: Study of Nucleus Vowel Duration and its Role in Prosody of Bangla. In: Proc. of Oriental COCOSDA 2007, Hanoi, Vietnam, pp. 181–184 (2007)
50. Scordilis, M.S., Gowdy, J.N.: Neural network based generation of fundamental frequency contours. In: Proc. IEEE Int. Conf. Acoust, Glasgow, Scotland, vol. 1, pp. 219–222 (May 1989)
51. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: A Standard For Labeling English Prosody. In: Proceedings of the 1992 International Conference on Spoken Language Processing, vol. 2, pp. 867–870 (1992)
52. Taylor, P.A.: A Phonetic Model of English Intonation, Ph.D. Dissertation, University of Edinburgh (1992)
53. Taylor, P.A.: The Tilt Intonation Model. In: Proceedings of ICSLP (1998); Taylor, P.A.: Analysis and synthesis of Intonation Using the Tilt Model. Journal of the Acoustical Society of America 107(3), 1697–1714 (2000)
54. Taylor, P.A., Isard, S.D.: A New Model Of Intonation For Use With Speech Recognition And Synthesis. In: International Conference on Spoken Language Processing, Banff, Canada (1992)
55. Yu, J., Tao, J.: The Pause Duration Prediction for Mandarin Text-to-Speech System. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEENLP-KE 2005 (2005)

Author Index